Unraveling the complexity of protein networks

Application in tumor immunity

Thesis for the degree of Philosophiae Doctor (PhD)

Trevor Clancy

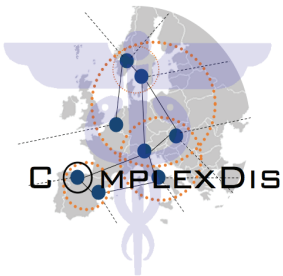
Department of Tumor Biology

Institute for Cancer Research

Oslo University Hospital

Faculty of Medicine

University of Oslo, Norway



Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1318

ISBN 978-82-8264-464-8

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Table of contents

Acknowledgements

First and foremost I want to thank my PhD supervisor Eivind Hovig. It has been an honor to be his student. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. Thank you for a great collaboration Eivind!

I would like to thank all of my co-authors and collaborators, who have helped me to implement this research. I have had the privilege of working with very talented colleagues from a wide range of backgrounds, and they have been instrumental in these projects. They have helped me to appreciate that bioinformatics is truly a multidisciplinary subject and teamwork with people from diverse scientific backgrounds can me most fruitful and stimulating. Many long vibrant discussions with molecular biologists like Timothy J. Lavelle often spur biological ideas. Another inspiration is working with the mathematician Einar Andreas Rødland, who has been a central guide to me in keeping me on the correct logical track, and in applying statistics to challenging biological questions in paper 3.

I would also like to thank Eirik Næss-Ulseth and all of my colleagues at the PubGene bioinformatics company. I was introduced to this company while completing my MSc in bioinformatics at Cranfield University, and then had the fortune of beginning my career as a Bioinformatician with PubGene. Their support and collaboration during my PhD period is appreciated immensely.

My family and friends deserve a lot of thanks in helping me complete this degree. My 3 brothers and 4 sisters in Ireland have encouraged me a great deal to achieve this level of education. Their moral support and encouragement throughout the years has been a great source of motivation for me. My sister Georgina has traveled to Oslo to represent them. She has always been a solid and generous voice of encouragement to me her little brother since I was a boy.

I have had the great fortune to meet very supportive friends and colleagues here in Oslo. They have been there always to offer support and advice throughout my years here and especially during the years of my PhD degree. In particular Marc, thank your for your advice, and encouragement and for being great buddy!

A special thanks to my girlfriend Christin! Thank you for putting up with me during this challenging period. You inspired me at the beginning to embark on this PhD degree to fulfill my ambition. Your inspiration was there at the end, supporting me in those crucial weeks before submission. I will always be grateful and happy for this. Thank you!

To my mother Margaret, to whom I owe everything in my life: There is no end to the gratitude I can express to you because there has been no end to your giving, sacrifice, love, and support throughout all my life. Doing all you have done alone is a remarkable achievement. You have achieved more than you can ever realize in your own life! This is an inspiration to me in my professional life. I am proud of you. If I can begin to repay you by making you a little proud with this PhD degree, I will be a very happy man. Thank you so much!

List of abbreviations

GWAS: *Genome Wide Association Studies*

GRNs: *Gene Regulatory Networks*

miRNAs: *microRNAs*

PPI: *Protein-Protein Interaction*

PTM: *Post Translational Modification*

IFN: *Interferon*

Th: *T-helper*

DC: *Dendritic cell*

TIL: *Tumor infiltrating lymphocyte*

NK: *Natural Killer*

CTL: *Cytotoxic T-cell*

IL: *Interleukin*

BDD: *Binary Decision Diagrams*

SAT: *Satisfiability*

KL: *Kullback Leibler*

List of papers

Papers in the thesis:

I. Pedicini M*, Barrenäs F*, **Clancy T***, Castiglione F, Hovig E, Kanduri K, Santoni D, Benson M. (2010). **Combining network modeling and gene expression microarray analysis to explore the dynamics of Th1 and Th2 cell regulation**. *PLoS Comput Biol*, Vol. 6, p. e1001032.

(*Shared first author)

II. **Clancy T**, Pedicini M, Castiglione F, Santoni D, Nygaard V, Lavelle TJ, Benson M, Hovig E. (2011). I**mmunological network signatures of cancer progression and survival**. *BMC Med Genomics*, Vol. 4, p. 28.

III. **Clancy T**, Rødland EA, Nygard S, Hovig E. (2011). **Predicting interactions between molecular machines from protein networks**. *Submitted manuscript*

Related papers by the author, not in the thesis:

1. Cekaite L, **Clancy T**, Sioud M. (2010). **Increased miR-21 expression during human monocyte differentiation into DCs**. *Front Biosci (Elite Ed)*, Vol. 2, pp. 818-828.

2. Agesen TH, Berg M, **Clancy T**, Thiis-Evensen E, Cekaite L, Lind GE, Nesland JM, Bakka A, Mala T, Hauss HJ, Fetveit T, Vatn MH, Hovig E, Nesbakken A, Lothe RA, Skotheim RI. (2011). **CLC and IFNAR1 are differentially expressed and a global immunity score is distinct between early- and late-onset colorectal cancer**. *Genes and immunity*.

Introduction

Molecular complexity of the cell

The classical paradigm of the cell

The cell is the unit of function, structure and phenotype in living systems[1]. Since its first observation by Hooke in 1665 [2], and then it's characterization in 1824 as the "fundamental element of organization" in an organism [3], the understanding of its complexity has evolved a great deal. In fact, the understanding of its complexity has progressed from being a "bag of enzymes" to a highly organized complex network of molecules[4]. The post-Mendelian theory was the dominant model to describe the phenotype behavior of the cell for over 100 years. This theoretical framework features a one-to-one correspondence of one gene to one function. Evidence for this one-to-one relationship was supported in 1923 based on causal alterations in one single protein, an enzyme of tyrosine metabolism, on the inherited disease alcaptonuria, published by Archibald Garrod [5]. What followed in the succeeding years was the "one-gene-one-enzyme" theory, a term coined in 1945 by Horrowitz [6]. His work, and that of his collaborators, Beadle and Tatum, on inherited defects in the mold *Nueorsposa crassa* [7], laid down the foundation for this long held concept that has served to explain cellular behavior for several succeeding decades.

Complexity & emergent properties of the cell

This once steadfast paradigm of one gene one protein functions has been replaced by the acceptance that the cell's molecules operate in a much more complex system of inter-dependent relationships. This system is governed by "non-linear" dynamics with emergent properties, *i.e.* the whole is greater than the sum of its isolated parts. This has become apparent in recent years through the many revelations brought to the fore

by the advent of high-throughput –omics technologies. For example, genome-wide associations studies (GWAS) has shown that much of the heritability of complex traits is apparently unexplained by initial GWAS, and this "hidden" component of complex disease still cannot be traced [8]. Historically, some have postulated that the number of possible molecular components in cellular organism is a determinant of complexity. In the case of genes for example, in 1964 Friedrich Vogel made a preliminary estimate of 6.7 million genes in the human genome[9]. He based this estimate on the accurate knowledge of the time, in addition to wildly incorrect assumptions that seemed justified at the time (such as all DNA being coding, and the average size of proteins). From the considerable volume of evidence accumulated in the subsequent years, we now estimate the number of genes in human approximately to be around 22,333[10], even lower than the 27,000 genes in the plant and model organism *Arabidopsis thaliana*[11] *(*hale cress*)*. Thus the number of molecular components in the cell of an organism has little relation to complexity and expression of phenotype behavior.

A network perspective of the cell

It could be argued that currently there is no solid theoretical framework that can effectively model the complex system of the cell, and how its large numbers of dimensions interact in a non-linear manner to produce a phenotype. However, that which has become transparent through the identification of the molecular components of the cell on the "–omics", or attempted comprehensive, scale, is the large network of molecular interactions, of various types, that occur within the cell[12]. Each of these networks could be considered to be subsystems of larger systems. These different systems of complex networks cooperate with each other, in a manner we do not fully understand, to manifest the phenotype of the cell.

The promise of computational and integrated bioinformatics approaches may help to build an accurate map of these interactions and to elucidate their mechanisms of function. This will at least lead to a better comprehension of complex cellular behavior. One widely used and rapidly evolving toolkit that can help us explore the dynamics and mechanisms of cellular complexity is the rapidly progressing field of network biology.

Molecular networks of the cell

Network interpretation of molecular interactions

Molecular cellular networks are the maps of the known components in a cell. These components include amino acids, nucleic acids, lipids, metabolites and metal compounds. They form a complex web of interactions that regulate biochemical homeostasis and determine the dynamic cellular response to external stimuli. High-throughput "–omics" technologies have been progressing rapidly in recent years to detect large sets of these molecular components [12]. Representation and analysis of cellular constituents through network principles is a promising and a popular analytical approach towards a deeper understanding of molecular mechanisms in a system-wide context [13,14]. The building and then deciphering of function from protein signaling networks also has great potential to aid discovery of new therapeutic intervention[15].

We can conceptually delineate three main classes of molecular cellular network being studied: metabolic, regulatory and signaling. Each of these classes of molecular networks represents physical binding interactions between molecular cellular components. These can be seen as somewhat conceptual classifications, as in reality these networks and their components work in an integrated fashion to respond to stimuli and confer the behavior of the cell.

Metabolic networks

Metabolic networks chart the interactions between all biochemical species in a cell[4,16]. In this class of molecular networks, the nodes are metabolites of chemical reactions in the cells, and the edges represent a description of the chemical reactions or enzymatic functions that alters the metabolite[4]. Although classic human metabolic pathways,

such as glycolysis and the urea cycle, have been studied for almost a hundred years, a complete human metabolic network does not exist. Limited maps in human of metabolic networks have been developed[17-20] and more extensive maps in prokaryotes are also in existence [18]. Some of these metabolic networks have been studied computationally in network models and simulations [21,22]. However, harnessing complete maps requires complete genomic knowledge and the complete acquisition of the functional relationships of all enzymatic proteins and metabolites. Although we fall significantly short of comprehensive knowledge of this biochemically-detailed class of network, there are interesting developments to construct and model metabolic networks. In human liver cells, for example, a genomic reconstruction of metabolic networks discerned metabolic states in at a large variety of physiological conditions[23]. The organization of metabolic networks have been shown to correspond to chemical properties which appear sensible for this organization[24]. The importance of the interplay between these small metabolites in regulating the activity of protein functions, through their integration with protein networks, has been reported recently[25]. Critical functions can now be placed on the metabolite-protein network, implicating the pathogenesis of many diseases and mechanistic action of various potential new drugs. This warrants future network studies of these biochemically-detailed networks, not in isolation, but as integrated cellular systems.

Gene regulatory networks

Gene regulatory networks (GRNs) are a class of molecular networks that are composed of transcriptional networks of gene regulation. In this class of molecular networks, nodes are protein transcription factors or a DNA regulatory sequence and the edges are directed to the binding of the transcription factor to the regulatory sequence. These networks are complex control systems that regulate the expression of

thousands of genes in any given process in life and are particularly important during the formation of life during development [26]. In recent years, there has been rapid developments to capture these relationships on a large scale in different organisms, namely the yeast one hybrid (Y1H) system [27,28], chromatin based ChiP-Seq[29] and ChiP-chip arrays[30]. Various computational models using networks have been developed and applied in recent years to analyze GRNs [31,32]. These methods range from the very first application of qualitative Boolean (logic based) networks in 1969 by Kaufmann[33], to continuous models that incorporate more dynamic and quantitative behavior of the gene expression using differential equations [33]. Incomplete knowledge and a mechanistic understanding of how gene regulation is governed in the cell limits the accurate modeling of GRNs. For example, there is increasing importance attributed to the role of micro RNAs (miRNAs) to regulate the mRNA expression levels in a cell[34]. For example, miRNAs are integral to the differentiation of monocytes, governing the expression of key protein networks that modulate this process[35]. The experimental strategies to capture this information are only now progressing on a large scale[34,36], and as a result a lot of these relationships are limited to computationally predicted targets of the miRNAs. Furthermore, the concept held for over 50 years of how a gene is regulated in a GRN, that of the Jacob/Monod lactose-operon explanation of a bacterial gene regulation circuit[37], is now known to be a much more integrated system involving the rich complexity of the entire cell[38]. Inherent in all aspects of GRNs are interconnected protein networks, involving interacting protein complexes or cellular machines, packaging the genome and organizing it in the nucleus[39] . This interaction network brings about a 3D conformation and compartmentalization compatible for a specific transcription factor program to render its gene expression program in the cell. Therefore, to truly

understand how gene expression is governed, we must capture complete information of the DNA and RNA regulatory elements that are integrated with protein networks in the cell to provide a platform for the cell to respond to stimuli and express its phenotype.

Protein networks

The third class, and the primary focus of this study, is that of protein networks. The underlying basis of these molecular networks consists of binary protein interactions. In these networks, the nodes represent proteins and the edges represent physical binding interactions between two proteins. These are also commonly phrased in the literature as protein-protein interactions or PPI networks. The classical and once dominant model for protein signaling networks is that of a canonical "pathway". This is a one-dimensional cascade consisting of tens of proteins, hierarchically organized, and independent from the rest of the protein network of the cell. The pathway model has been useful as a tool to explain the properties of some cellular functions, and pathways have been catalogued in many useful databases [18,40,41]. However, the pathway paradigm is a limited and linear conceptual framework to understand both normal and disease cellular behavior. Their limitedness in scope and coverage across the many interconnected cellular processes resulting in them missing many important interacting protein pairs, make these resources inconistent[42]. New models and analyses of large-scale protein networks are evolving to respond to emerging high-throughput technologies that allow for a very alternative view of signal transduction[43,44]. It is increasingly apparent that these large-scale screens and their accompanying network analyses are taking precedence to study cell signaling[44]. The acquisition and analysis of protein interactions is critical to gain a systems level understanding of the cellular complexity[45,46]

Protein networks & crosstalk mechanisms

The study of the cell and its relationship to diseases such as cancer is likely to benefit a great deal from a global network understanding of how information signals are propagated in the cell. Such a global network view of the cell can facilitate important studies that characterize crosstalk in protein signals in disease states, like for example that of the EGFR and insulin receptor pathways[47] or between CDK8 of the mediator complex and β-catenin activity in colorectal cancer[48]. Crosstalk mechanisms have not been studied as thoroughly as linear signaling pathways, and network biology is opening up to their discovery and characterization[49]. Characterizing crosstalk signals in molecular networks will be crucial to understand pathogenesis, particularly so in cancer[50] and the immune system [49]. In many cases, this will lead to beneficial clinical outcome. For example, a protein network approach has recently lead to an improved understanding of the resistance of melanoma cells to the BRAF kinase inhibitors, demonstrating the importance of pathway crosstalk signaling in drug inhibition[51]

Biochemical modifications in protein networks

The complex phenotype of a cell may also be seen as a function of the different biochemical states a protein may be in, and also as the complex network of interactions between the species these states create. Protein networks, in the true reality of the living cell, are not the static structures as we see them corresponding to function. Rather, protein networks exist as cooperative systems, communicating through various different biochemical mechanisms to propagate signals. These are collectively termed post-translational modifications, such as phosphorylation and ubiqutination. In many cases, they effectively confer a different species of function onto a protein that will then determine its fate and pattern of further interactions.

These chemical modifications to proteins are also being populated into various databases[52-54] and are increasingly used in network models. There are studies that have achieved positive results in using protein networks to predict phophorylation sites, an improvement on the standard sequence based predictions for these sites[55,56]. Another study has manually curated ubiquitin posttranslational modifications (PTM) on to binary interactome data and has computationally identified high-confidence interaction signals[57]. For the most part, this level of information is absent in network analysis and large-scale screens and their accompanying networks approaches, because of the sparse amount of experimental verification of sites of PTMs. This is the main current drawback on using large-scale protein networks approaches compared to the canonical pathway model. Although adding additional layers of complexity, as this information is populated into rapidly growing databases or harnessed through advancing proteomics approaches[58,59], applied to specific cellular processes such as phosphorylation during mitosis[58] or apoptosis[60]. The integration of these upcoming resources and their analysis in protein networks will be a powerful area of future cellular network biology research.

The acquisition and analysis of protein networks

Community efforts to organize & structure protein networks

To gain an understanding of the complex processes occurring in the cell through its protein networks, it is crucial that all protein interactions are eventually identified and adequately organized. It is estimated that most of the binary protein interactions in the human protein network remain experimentally undiscovered [61]. There have been many efforts in recent years to experimentally harvest these protein interactions using high-throughput experimental procedures in human[62-64] and many model organisms, such as yeast [65-67]. There are ongoing research efforts to improve the quality of these binary interactions to produce high-confidence connections [61,68]. Efforts to study the quality of interaction networks have reported extensive incompleteness and noise[69,70].

Independent efforts are progressing continuously to build comprehensive protein network databases[71-79]. With the growing number of protein interactions being catalogued, it is essential to use organized relationships of interactions in a consolidated and non-redundant manner[80]. Important community efforts are underway to achieve this [81-83] and their results will be central to any application to study the protein complexity of cells.

Data mining for protein networks

Although protein interaction information resources are continuously expanding, they are still very much incomplete [61,84]. For that reason, prediction methods hold great importance to acquire a more complete perspective of cellular complexity and to infer various biological relationships of complex phenotypes. One common strategy to predict protein interactions is to use the conserved sequences, or functional domain sequences of known protein interactions, *i.e.* their binding interfaces, to infer putative

interactions [85-105]. These sequence and protein domain family methods have resulted in various levels of success. Another approach is to use the genome features of protein pairs to predict interactions. One such experimentally validated effort to computationally predict interactions has demonstrated that using a Bayesian analysis applied to genomic features is very promising in discovering novel protein interactions[106]. The STRING database integrates multiple sequence, literature and experimental parameters to predict interactions[107], and its comprehensiveness has made it a popular source for protein interactions.

Text-mining for protein networks

An alternative approach to capture the enormous scale of protein interactions in the cell is to use automatic extraction of these relationships, from the ever-expanding 20 million-plus articles in Medline [108]. Indeed, in general terms literature mining has now distinguished itself as a viable method to capture and organize many types of biologically relevant information [109-111]. With respect to protein networks, literature mining approaches have been used for over 10 years to extract protein interactions from the Medline database. First, based on the simple rules of co-occurrence of two proteins mentions in an article's abstract[112,113], and later progressed to more elaborate procedures that incorporate machine learning [114,115], Bayesian inference[116], linguistic[117,118] and ontology [119,120] based approaches. Both these automated literature mining methods, and the time consuming process of manually reading and curating the literature for protein interactions[121], are both error prone and replete with biases. One of the most often highlighted biases is their containing more interactions for well-studied biomedical concepts[61,121].

Boolean modeling of protein networks

All control of the molecular components in gene regulatory, protein, and metabolic networks are governed by a variety of biochemical mechanisms, with inputs from other network components that act additively or synergistically on the molecule in question. At present, most knowledge we have about protein networks is mainly qualitative in nature. Because our existing knowledge of complex protein networks is based on discrete qualitative values, Boolean models are appropriate models to analyze their behavior. Boolean networks, as applied to signaling in protein networks, are based on the assumption that binary "on" or "off" states functioning in discrete time steps and describe important aspects of outcome of the network. They have traditionally being used in the modeling of GRNs. The simplest dynamic models applied were developed for small random networks of transcriptional regulation in the 1960's by Stuart Kauffman[122,123]. Boolean networks have been limited to small networks in the past as a given network of *n* genes or proteins, there are a total there are 2^n possible different phenotype states. This makes the updating of all possible states in real cellular networks inexorably large, and difficult to model in reality. The succession of states with time is monitored and a record is kept of which states are reached at each update. Some states may never be reached. The goal is find attractors: these are states or series of states that once reached, remain stable. The attractors can be synonymous with phenotype behavior of a cell as measured through experimentation, such as a gene expression outcome or a signature cytokine released by the cell in question. Kaufman considered each attractor as a stable differentiated state of the cell in 1969 when he first devised the approach[123]. In later years, he demonstrated that the number of differentiated cell types predicted by this model corresponds well with the current experimental knowledge[124]. The majority of studies

applying Boolean approach have dealt with GRNs, most of which were small in size (ca. 10 nodes). They have however begun to be used for logical analyses of signaling networks. One of the earliest examples of this trend was in 1999, using a small simplistic Boolean model governing the signal transduction of effector T-cell activation was formulated[125].

Alternative logic models to analyze protein networks

In many cases the relationships in protein networks are too complex to be captured with simple Boolean logic, and therefore more general models have been developed. These models are still discrete model types, in addition to the types of networks previously analyzed Boolean networks, are so-called logical models[126], Petri nets[127] and agent-based models[128]. It is possible that these approaches could be better performing solutions to elucidate the cellular mechanisms in complex molecular networks, such as that of T helper cell differentiation.

Structural properties of cellular protein networks

The application of network theory to cell biology has fundamentally altered our understanding and appreciation of the complexity of the cell[4,129-131]. There is no formal definition of the complexity of a cell, but networks provide us with an adequate set of tools to explore the relationships between the extraordinary high numbers of molecular components in the cell. The advent of these tools and network discoveries in recent years has provided the field of network biology with significant advancements. It has been observed, for example, that the topological properties distinguish real cellular networks from random networks[4,132]. Some of the topological properties have received very noticeable attention in the literature in the past decade. For example, the distribution of degree (the number of interactions per protein) in

cellular networks is often claimed to follow a power-law distribution[4,133]. This property of protein networks and metabolic networks has been found in all organisms where data exists, from yeast to human[134]. This, and its correlated features in protein networks arise important biological questions, the solutions to which may help to unravel the complexity of cellular networks and lead us to a perception of the functional organization.

Tools for Network visualization

Extracting relevant information from this huge amount of data becoming available for cellular network analysis requires dedicated tools. Such analysis of visual, topological and dynamic properties of cellular networks is a highly active area of research and development. Several very effective tools have been developed to address the need for network analysis. Some of these focus on the simple visualization aspects for data exploration and integration tasks[135-140]. Other tools have been developed to offer more sophisticated analysis pipelines for integrating multiple datasets, for cluster analysis and to investigate the topological features of the network.[136,140-142]. Other tools have been developed for the dynamic analysis and implementation of systems biology models[143]. The recent rapid advanced in these tools have allowed typically large networks comprising several thousands of proteins and their interactions to be analyzed, efficiently and seamlessly.

Inter-connectivity of protein networks in disease

Network approaches offer an improved understanding of the relationship between the genes implicated in diseases [15,144-147] and may be a valuable resource to find candidate disease genes[148]. It has been reported that the Mendelian component of complex diseases, such as for example breast cancer, represent less than 30% of its incidence [149]. In the particular case of breast cancer and the BRCA1 and BRCA2 genes, it is a mere 5% of all cases[149]. Furthermore, the recent results of the many GWAS undertaken in recent years have shown that a large amount of disease-causing genes are yet to be accounted for [8]. To explain the missing causal factors of complex disease, it is suggested future investigations should focus not on the genes in and of themselves, but rather on the effect of the interaction of their protein products and perturbation of the cells protein networks [14,145,146]. For diseases of simple Mendelian inheritance, it is suggested from their expression patterns that they have central importance in protein networks [150]. In contrast to those arguments, the majority of Mendelian disease genes show no tendency to have high connectivity in protein networks, and their expression pattern indicate that they are localized in the functional peripheries of the network[130,151]. This makes sense in the light of most highly connected protein being those of essential genes[130,151], and therefore their mutations would be deleterious during fetal development. Looking at these disease causing mutations from a protein structure perspective suggests that approximately 4% of single-gene disease mutations have an effect on the binding interfaces between protein interacting pairs[152]. Interestingly, there is a high level of disease gene clustering in protein networks[151] detected, despite our current very limited knowledge

of protein networks[84,153]. This high degree of clustering of disease proteins in networks of the cell is the key element to future discovery and research of disease causing factors. The interconnectedness of disease proteins in communities or modules of interacting proteins may well be a source of the pathogenic phenotype. These disease proteins with high clustering possibly correspond with functional modules or protein complexes that are important to normal cellular function. This local clustering is important, as interactors of the disease gene, and not necessarily the disease gene itself, has important biochemical implications on a cellular process[154]. It is therefore proposed that in concert with genetic variation, protein interactions and the networks in which they operate are central to the pathogenesis of complex diseases and therefore a fruitful source of future disease gene discovery. This is shown to be increasingly the trend in light of the capturing of disease-associated protein network modules in large-scale screens of protein networks, in complex diseases ranging from autism [155] to Alzheimer's [156] disease and heart disease[157].

Protein networks & cancer

It has been proposed that an analysis of the key properties of proteins implicated in cancer in protein networks will guide the discovery of candidate targets for therapeutic intervention[158,159]. Contrary to the location of inherited disease genes in protein networks, the somatically mutated genes in cancer have a tendency to be found as central hubs in protein networks[130,151]. This notion of cancer genes having central roles in protein networks was also put forward by evidence of differentially expressed genes up-regulated in lung squamous cell carcinomas having significantly higher number of interactions partners in the human protein network [160]. Similarly, an investigation of the known tumor suppressors and oncogene proteins [161] indicated that they have double the number of interaction partners when compared to non-cancer

proteins [162]. However, this evidence may be result of the bias in cancer proteins being studied far more often, and therefore over-represented in protein networks. For example, an analysis of 29 cancer differential gene expression studies against 22 different metrics of network properties indicated no strong evidence for a large number of highly connected proteins, but a higher degree of interconnected modules or groups clustering among cancer proteins[163] was found. It is clear that somatic mutations are frequently involved in functional canonical pathways, as was revealed by the DNA sequencing of 623 genes with known or potential relationships to cancer[164]. Furthermore, a network strategy based on the analysis of mutations within network modules in several cancers identified rare cancer driver mutations involved in key cancer pathways. In that study, the genes identified do not play a central role in the pathways, but rather contribute greatly to a more refined tuning of function of these modules through possible crosstalk mechanisms[165]. It is clear that a focus on a modular analysis of groups of interacting proteins that correspond to protein complexes or functional networks, rather than a linear pathway analyses, are proving enormously effective in prioritizing the molecular factors of cancer progression[166,167]. A modular analysis of cancer protein networks has proven to unravel complex intertwined oncogene RAS pathways in cancer cell lines, whose functions are connected to processes that mediate sensitivity to drug response[168]. The observation of the phenomenon of crosstalk in cancer protein network modules was important in that particular study, and strategies are being developed to capture proteins and protein network modules that cross talk with each other[169]. The phenomenon of protein network modules cooperating with each other to confer the phenotype in a cell has been modeled using Boolean logic to indentify protein network signatures that have significance to clinical and biological outcome [170]. These and many other related

studies highlight the utility of dissecting protein networks to help us understand the cells complex phenotype.

Complex networks of tumor immunity

Molecular complexity of the immune system

Similarly to tumor cells, there is a substantial amount of signal transduction, with frequent and diverse crosstalk and sharing of protein components, among signaling protein networks in immune cells[49,171]. For the immune system, this is primarily cytokine-mediated cellular communication[171,172]. For example, the IFN-γ protein network is implicated in crosstalk to multiple signaling cascades, other than its well-characterized regulation of activation of the STAT1 gene expression program. For now, the crosstalk behavior of IFN-γ is not comprehensively understood[173]. This makes IFN-γ protein networks an exemplary target for discovery of immune signaling in complex protein networks. The transduction of information signals through these very complex protein networks makes it a daunting task to elucidate biological meaning, not least for immunologists who treat signal transduction in networks as linear canonical pathways[174]. The traditional approach to understand the immune system by immunologists has involved deconvoluting the complex heterogeneity of immune cells with flow cytometry, using combinations of markers to define signatures that represent specific lineages, differentiation states, and functions. This strategy of studying complex immune phenotypes on a single protein basis is easily measured, visualized and interpreted. However, to capture a true understanding of immune phenotypes involves identifying dynamic changes distributed across complex networks of proteins. This is far more challenging. Furthermore, the current biological models of complex human immune system signaling are based on an over-reliance on the mouse model, which has been disappointing in the study of human immunological diseases [175]. The mouse has 65 million years of evolutionary distance from human and

in research environments is subject to a skewed immunological profile due to an over-abundance of homozygous recessive mutations, caused by excessive inbreeding [176].

Global approaches to immunological discovery

New strategies are currently evolving to address these challenges, and now a global view of human immune signaling is emerging[177]. These strategies are computational in nature and are progressing from initial efforts in computational immunology to building of immune databases [178,179] to computational network modeling approaches[180-182]. A global approach to capture modules or communities of proteins has been applied, and has successfully identified canonical pathways implicated in the mRNA expression changes in patient blood during the immune response to lupus [183]. Another systems approach developed a vaccine-behavior prediction method that performed with very high accuracy[184]. These and other studies signify an emerging trend of applying computational methodologies designed to support a systems-scale analysis of the immune system[177].

The future progression of these approaches is very much dependent on accurately and comprehensively mining and capturing protein network modules that are significant for the immune response. Prior to the extraction of protein network signals from patient samples, there is the seemingly difficult challenge of clarifying the definition of an immune gene. There are several international efforts underway to make these definitions and catalogue immune genes into databases [185-187]. The methodological development that comprises the strategy outlined in **Paper II** of this thesis, describes a great deal of disparity and disagreement in these immune gene databases[188]. In addition, that study implicated a great number of genes associated to the immune response, as yet uncharted by contemporary immune gene databases.

Protein network approaches & the immune response

There is strong evidence linking genes of immunological diseases to highly interconnected modules or clusters in protein networks. This was shown to be the case recently when analyzing 150 different GWAS loci tightly associated to immune diseases and demonstrating an abundance of highly connected protein interactions between the protein products of genes in these loci[189]. Faced with the complexity of immune cell signaling[49] and the plethora of possible cytokine interactions in tissue[190], network approaches to dissect the functional association from protein networks in immune phenotypes is warranted. Many such research projects have already begun in this direction. For example, a microarray-based study in blood leukocytes, stimulated by bacterial toxins, applied a network analytical approach to identify novel protein network modules that correspond to the molecular machinery that responds to inflammation and a septic shock, during the innate immune response[191]. Another recent study used a systematic experimental approach to treat macrophage cells as a conceptual "black box" for deduction of the properties of the protein-signaling network upon stimulation of cell receptors by six different "input" cytokines [192]. Their results suggest that the complex nonlinear networks in normal immune cells have a limited number of "outputs" (secreted cytokines), from the multitude of possible outputs. Therefore, complex protein networks are tightly regulated and controlled in the normal cell. Understanding aspects of this regulation would require completed large-scale protein network screens of immune cells. This has been achieved recently in B-cells using co-immunoprecipitation experiments and subsequent assembly of B-cell specific protein networks. Coupled with algorithms to interrogate this network, a valuable resources was created to allow for an elucidation of the phenotypes that control the complexity of B-cell regulation [193]. In that particular study, two novel

master regulators of the humoral immune response were discovered[193]. Furthermore, a recent effort that is seminal to protein network analysis at large, and specifically to tumor immunity, is the global protein network screening followed by a functional network analysis identified for IFN-γ signaling[194]. A pathogenic role of T-helper 1 (Th1) cells and IFN-γ in autoimmune diseases and cancer raises the question of mechanisms by which IFN-γ contributes to pathogenesis, which could be answered by network analysis of this resource[194] . The interactions between such inflammatory cytokines are currently being scrutinized for their involvement in modulating growth of invasive tumor cells[195], and cancer stem cells in the tumor microenvironment[196]. Attempts of transforming these developments of protein interaction network analysis are now ongoing in the clinical arena, where gene expression analysis of circulating immune cells linked to their protein interaction has been shown to identify pathogenic network signatures[197].

Tumor immunosurveillance: a brief historical perspective

During the 1700s, it was recorded that certain infectious diseases could have a beneficial therapeutic effect on malignant tumors[198]. This beneficial effect and regression of tumors was observed in certain cancer patients that developed bacterial infections[198]. The German pathologist and father of cellular pathology, Rudolf Virchow, documented influential observations in 1863 of the "lymphoreticular infiltrate", linking the origin of cancer to sites of chronic inflammation [199]. Later, the American physician William B. Coley in the 1890's began to pursue the relationship of immunity and cancer, when he noted that some sarcoma patients who had severe post-operation infections at the tumor site, underwent spontaneous and sustained tumor regression[200]. He, among others during that period, followed up with very

controversial experiments, with beneficial clinical outcomes, involving the deliberate induction of erysipelas (*Streptococcus pyogenes*) in cancer patients, with the intention of bringing their malignancies under control[200,201]. The concept that the immune system could eliminate primary tumors naturally, in the absence of external therapeutic intervention was first proposed in 1909 by Ehrlich[202]. This has been a point of heated debate and was not resolved until the acquisition of solid molecular evidence in recent years[203].

Thomas and Burnet coined the term "immunosurveillance" for this hypothesis, and developed the concept further during the 1960s [204-206]. Jonas Salk wrote a very forward-thinking essay on this topic in 1969, where he proposed that chronic infections, allograft rejections, autoimmune disorders and cancers belong to a common phenomenon known as "delayed allergic reaction"[207]. This line of research began to develop in an era when experimental models were finally becoming available to test the immunoesurveillance hypotheses. However, using mutated mice models that rendered an inactive immune system in the animals (nude mice), results were derived that contradicted the hypothesis. There was clear evidence indicating that the nude mice did not develop spontaneous tumors[208 209]. When no difference in primary tumor development was found between these mice and wild-type mice, the immunosurveillance concept was largely abandoned.

The broad acceptance of the phenomenon did not take hold until as late as the 1980's, when it became apparent that nude mice models were immune-compromised, but not completely immune-deficient. The nude mice model used in previous studies did not completely lack functional T cells[210]. In the years that followed, the proteins responsible for immune mediated tumor suppression began to be identified. The pro-

inflammatory cytokine IL-2 was shown to clearly contribute to tumor regression in metastatic melanomas[211], and IFN-γ prevented tumor formation in mice[212,213]. However, IFN-γ was also shown to collaborate in selecting for tumor cells with reduced immunogenicity, leading to malignant cells that are more capable of surviving against immune attack[214]. This explained possibly why immune competent individuals still develop cancer. These paradoxical roles of the immune system on the development of a tumor, prompted a re-definition of the cancer immunosurveillance hypothesis in recent years into cancer "immunoediting"[203,215,216]. This now accepted phenomenon has taught us that the immune system plays a dual role in response to a tumor. It can suppress tumor growth by killing cancer cells or by inhibiting outgrowth. It also can promote tumor progression by selecting for invasive tumor cells or by establishing favorable conditions within the tumor microenvironment. That which is not understood entirely, are the complex protein networks that mediate this process within and between cells in the tumor microenvironment.

Global approaches to tumor immunity

With the advent of high-throughput technologies and more robust experimental models in immunology, there has been a rapid increase in the number of identified molecular players implicated in the tumor immune response[172,217-219]. This increase in evidence for relevant immune factors comes with increased complexity in the networks of relationships between these molecular players. This corresponds with the now accepted paradoxical biological and clinical outcomes that the immune system has on a tumor[220]. In turn, these revelations are coupled with the increasing trend of large-scale studies to capture the complete maps of protein interactions that regulate the major players, such as that of the recent screening of IFN-γ protein network[194].

This will offer us a resource to an increasingly detailed perspective of IFN-γ and its mechanisms of complex crosstalk in protein networks[173]. With the continuous increase in such large-scale screens, we are soon on the road to the discovery of a complete list of molecular players, and a map of the complex networks that contribute to the relationship between immunity and cancer. With this comes the necessity to develop computational strategies to mine, organize and decipher the complex protein networks that govern the balance between immune tolerance, promotion or rejection of a malignant tumor.

Protein networks & immunity in the tumor microenvironment

So, it is now well established that cancer is an inflammatory disease[217-219,221,222] and that immune cells are recruited to and infiltrate into the microenvironment of a tumor[223,224]. There is increasing amount of recent evidence suggesting that some patients with cancer can mount an antitumor immune response that has the potential to control or eliminate cancer[223]. Numerous reports have appeared in the literature confirming that the infiltration of immune cells into a tumor plays a crucial role on the survival of patients. In these patients, an immune response signature (*i.e. a* community of genes) has been described, that is associated with improved outcomes in several tumor types. This has been reported for colorectal cancers [225], follicular lymhpomas[226], melanomas[227] , and ovarian cancers [228]. T-cell environments that secrete cytokines such as IFN-γ generate acute inflammation that results in expansion of natural killer (NK), cytotoxic T cells (CTLs), M1-macrophages, tumor destruction, and the potential control or even elimination of cancer[214,216]. These signatures are associated with Th1 immunity and acute inflammation, similar to graft rejection. In more aggressive malignancies, immunosuppressive environments are described that

promote tumor proliferation, and protect the tumor from immune attack or clinical interventions[229,230]. This is inflammation of a different type, a chronic inflammation characterized by the IL-6 cytokine[217,231]. A term often used to describe this phenotype is "smoldering inflammation"[217], and is an environment that is similar to wound healing mediated by Th2 cells[232]. It produces the cytokines IL-4 and IL-13, TGF-β, which suppresses anti-tumor Th1 immune responses, and EGFR ligands, which promote tumor growth and metastasis[232,233].

Regardless of the direction of the immune response toward a tumor, the phenotype outcome in the microenvironment is mediated by complex protein networks that promote inflammation in cancer development[229,234]. The type of protein interactions presented to immune cells will then affect the type and nature of protein interactions, and thereby the immune response by those cells. These protein networks are both intrinsic in, and extrinsic to, all cells in the tumor microenvironment: normal, fibroblasts, sentinel-immune, endothelial, tumor, infiltrated immune cells, etc. During cancer progression, dynamic protein interactions occur between tumor cells and host immune cells that may function to either stimulate or inhibit cancer growth. These protein interactions also facilitate various cells to communicate with other cells in the local microenvironment, by secreting various protein-interacting cytokines and growth factors, or hosting these factors on their cellular membranes. These complex immune phenotypes are a challenge to capture from the tumor microenvironment. Methodologies to quantify the immune phenotype in tumors were developed in **Paper II** of this thesis, to address these challenges[188].

The entire complement of these factors is called the "secretome". This term was coined by a review in 2009 that summarized evidence suggesting that the secretory

phenotype of senescent cells fuels inflammatory responses that in turn recruit immune cells to create immune clearance phenotypes[235]. Secreted cytokines and chemokines are manifested by all cells and is a process that becomes increasingly complex during cancer progression. This is especially so in the immune clearance of oncogene induced senescent cell in cancer[236]. The delicate balance in the tumor microenvironment, switching between immune-surveillance, -tolerance, or –escape, is dependant of the nature of activation of the adaptive immune system.

Th cells & protein networks: inflammatory switches

Disrupted T helper cell (Th) responses can cause a range of diseases, including cancer. The Th-cell responses are coordinated through distinct functional protein networks, governed by distinct programs of transcription factors that ultimately have distinct consequences for a malignant tumor. They recruit to the microenvironment, and activate other immune cells to respond to a progressing tumor. These activated and recruited cells include B cells, NK cells, macrophages, mast cells, neutrophils, eosinophils and basophils. Th-cells regulate these immune responses via the production of specific cytokines, which act as messengers to instruct other cells of the immune system. There are currently four defined $CD4^+$ Th-cell subsets: Th1, Th2, Th17 and T_{reg} cells[237,238]. Th1 is characterized by the stable expression of the cytokine IFN-γ, and coordinates tumor-killing responses. Conversely, Th2 is characterized by the stable expression of IL-4 and coordinates metastatic tumor-promoting responses. The classic paradigm from its conception in 1986[239], was that the Th lineage was thought to exist strictly in a dichotomy between the Th1 and Th2[237,240] cell lineages, *i.e*. Th1 and Th2 were stable states expressing a clearly defined output of cytokines, and were antagonistic regulators to each other. For some time, Th1 and Th2 were

considered to be the only types of $CD4^+$ effector responses. However, the latest experimental reports contradict this dichotomy. In fact, it has become apparent that $CD4^+$ T cells undergo a complex process of differentiation enacted through complex signaling networks. Th cells differentiate not only into stable lineages of Th1 and Th2, but also into two other major lineages: Th17 and T_{reg} cells[237].

This process is dependent on the functional interaction stimuli received by the naïve $CD4^+$ T cell, the pattern of cytokine secretion of the various lineages and the protein signaling cascade that leads to a defined expression of specific transcription factors. Th cells from different lineages secrete their characteristic cytokines, resulting in a much greater degree of heterogeneity of the Th cell population than was originally thought possible. In addition, the pattern of cytokine secretion switches from one lineage to another under different phenotype cues from the tumor microenvironment. This indicates that Th cells exhibit great plasticity in their lineage commitment, which has important implications for the fate of a developing tumor[238,241,242]. This plastic process of cellular differentiation of Th cells is akin to "decision-making" by the naïve $CD4^+$ T cell precursor cell[243-245]. This process is governed by complex, yet orchestrated, protein networks, which have clearly defined cytokine-inputs and cytokine-outputs. Thus, this can be seen as an ideal protein network cascade that can be analyzed using computational networks modes as is done in **Paper I** of this thesis[246], and other network studies of Th cell regulation[247,248]

Cellular machinery in protein networks

Protein networks & networks of molecular machines

Proteins rarely function alone [249-251]. Therefore a sensible interpretation of complex protein networks in the cell will require an analysis of their actual mode of function. In physiologically relevant states, the peptide sequences of proteins are transformed into three-dimensional structures, which bind stoichiometrically to other peptide units at the same time and cellular location, to form a quaternary structure, *i.e.* the "molecular machine" or protein complex[252]. These are the actual functional structures that carry out most processes in the cell, such as the ribosome or membrane synapse of a T cell. It is becoming increasingly apparent that new approaches are needed to transform the rich information in protein networks into knowledge of these molecular machines[251]. In order to achieve a complete understanding of cellular complexity, the detailed mapping and structural analysis of molecular machines in the cell needs to be carried out[251]. This will entail massive efforts in the identification, isolation, structural characterization and mechanistic analysis of these machines[252,253]. Computational prediction methods may be valuable in this endeavor[254,255]. Traditionally, many protein network studies have treated the fundamental unit of function in cells as the proteins. In **Paper III**, using the most comprehensive databases available on yeast[256,257] and human[258] protein complexes, the complexes are treated as nodes in cellular networks, and "higher-order" interactions are predicted between these units.

Permanent & transient protein interactions

The binding affinities of the protein interactions are important features of protein complexes. Protein interaction can be categorized into two types based on their binding affinities to each other. Permanent protein interactions usually form stable

structures together in molecular machines. Transient interactions are less stable. They associate and disassociate from each other quickly and temporarily[259,260]. In the current status of knowledge in protein network databases, interactions are not annotated into the two categories. In the cell, there is a continuum existing between transient short-lived interactions and permanent interactions found in stable functional molecular machines, making it difficult to resolve which interactions are corresponding to protein complex formation, from the short-lived transient interactions[260]. These dynamics very much depend on the physiological conditions of the cell[261]. The protein networks in existence today underlie both of these inter-mixed categories of protein interactions. Many proteins are involved in more than one protein complex and binary interaction. These complex features need to be characterized in terms of their detailed structures and kinetic mechanisms in order resolve completely the complexity of protein networks.

Methods to identify transient interactions between proteins are being improved and facilitated constantly by the accumulation of protein network data from proteomics and structural biology[262]. Continuous development of technologies that are fine-tuned for the detection of weak protein interactions and their structural features[263] will complement many computational approaches to understand their role in complex protein networks. The incorporation of the detailed biochemical and structural information of molecular machines can convert an entangled complex network of binary protein interactions into accurate biological models. The growth of these protein complex structures and their dynamic properties will improve on computational procedures to predict novel relationships in complex protein networks[252,254], such as that reported in **Paper III**, *i. e.* to predict the higher order complex-complex interactions.

Protein complex databases and proteome-wide maps

There are increasing large-scale efforts to build an accurate perspective of the protein complex repertoire of cells. Such proteome-wide interactome maps are emerging over the past ten years, offering increasingly comprehensive data on binary protein interactions and protein complexes. These exist for many organisms, including *E. coli*[264], yeast[65,265] and human[63,64]. The next generation of these proteome-wide maps are adapting to acquire, more comprehensively, experimental data to provide evidence for the interactions between stable complexes[266,267]. In these next generation proteomic studies, large-scale laborious efforts using a wide range of proteomics toolkits across several laboratories have attempted to capture experimentally the global organization or "complexome" of human [267] and *Arabidopsis*[266] protein complexes.

These developments are complementary to the goal of computationally inferring this knowledge, from manually curated complexes and binary protein interaction maps, such as that attempted in **Paper III**. These inter-complex interactions are difficult to experimentally detect precisely, without erroneous inferences and assumptions. They therefore necessitate some degree of statistical inference to rank relevant inter-complex protein interactions. With the advent of a growing number of these experimental resources of protein complexes, comes an increase in specialized databases to store and organize this information. Outside of the protein network databases mentioned previously, there are dedicated databases of manually curated protein complexes for various organisms that are proving to be very useful for proteomic studies[71,256-258,268]

Aims of the study

This study had two main objectives: (1) To develop computational strategies to provide a broadened understanding of complex protein networks in the cell, with particular application to immune cells critical to cancer progression. (2) To treat protein complexes as the functional units in the complex protein networks of the cell and predict networks of interactions between these units. Thereby, attempting to resolve the complexity of the cell by treating it as a functional network of interacting molecular machines.

The study attempted to achieve these goals through developing and applying logical protein network models in a cellular phenotype within a clearly defined immune cell, *i.e*. Th cell differentiation. In addition, strategies were devised to quantify the immunological phenotype in cells of complex tissues, specifically tumors, and to allow the identification of immune related protein networks linked to cancer progression. The purpose of these undertakings was to achieve an improved understanding of cellular organization and protein network complexity, making an abstraction in cellular protein networks to their core biochemical components, *i.e*. molecular machines and predict interactions between these protein machines.

The following are the specific aims related to each of the three papers:

I. To apply a Boolean model regulatory network, based on an extended signal transduction network and directed protein interactions implicated in Th cell differentiation. The intent was to improve our understanding of the complex networks implicated in the differentiation of T helper (Th) cells into its regulatory lineages. Further to this the goal was to examine whether this

network approach could assess the Th1/Th2 paradigm with respect to its compatibility with a counter regulatory role of the Th1/Th2 lineages.

II. To develop an integrative computational approach that quantifies the immunological relevance of all genes in the human genome. The specific goal of which was to capture and quantify the signatures implicated in the immune response during cancer progression, from large-scale measurement of genes detected in high-throughput experiments. Further to this, the goal was to map this quantitative immune information to protein networks, to provide insights to further hypotheses on the mechanism of the tumor immune response during cancer progression

III. To investigate globally the biochemical mechanisms in protein networks by abstracting protein networks the level of molecular machines (protein complexes). Interactions between these molecular machines were predicted, as well as the use of these interactions as representative model of cellular networks

Summary of the Papers

Paper I

In this paper, a network analysis of Th cell subsets, namely Th1 and Th2 cells, was carried out. This allowed an assessment of the long held conception that these Th subsets counter-regulate each other to orchestrate inflammatory responses. Understanding these dynamics would allow for future studies on how alterations in their balance may result in different inflammatory and autoimmune diseases. This is particularly relevant in cancer, where the Th status of the tumor microenvironment is critical to clinical outcome. This paradigm of the Th1/Th2 balance has been challenged by recent clinical and experimental evidence. There are a large number of genes involved in the signaling cascades and regulation of Th cell differentiation, and therefore an assessment of the Th1/Th2 paradigm by modeling or experimentation has been very challenging. However, a recently developed novel algorithm caters for the analysis of much larger model Boolean logic networks[269]. Using this, combined with other levels of computational analysis, such as *in silico* knockouts, and gene expression microarray data from human T cells, we examined if a network model was compatible with a counter-regulatory role of Th1 and Th2 cells.

We constructed a directed network (including aspects of activation and inhibition) of genes regulating Th1 and Th2 cells through a combination of literature mining and manual curation. Application of the Boolean model on this network identified four attractors in the network, three of which included genes that corresponded to Th0, Th1 and Th2 cells. The fourth attractor contained a mixture of Th1 and Th2 genes. It was found that neither the *in silico* knockouts of the Th cell attractor genes nor the gene expression microarray data from patients with immunological disorders and

healthy subjects supported a counter-regulatory role of Th1 and Th2 cells. These attractors were identified along with an additional attractor we named ThX. In some respects this may reflect the current biological understanding of rapidly changing Th1/Th2 paradigm. Overall, this paper indicated that combined network modeling, *in silico* knockouts and gene expression microarray analyses, is a tractable approach to unravel the complex signaling and regulatory networks of cells.

PAPER II

It is clear from some of the outcomes in **Paper I**, that there is great deal of unresolved complexity, most of which is not yet understood or acquired, associated to Th cell populations. Recent evidence suggests that the immune component of the tissue to which the Th and other effector cells of the immune system migrates to determines their final differentiation state[270] Therefore, identifying the immune molecular components and quantifying their signal in complex tissue, not least in tumors, would be very beneficial for future studies in understanding this complexity. This was the motivation behind **Paper II** of this thesis.

The immune signal in complex tissue, not least in tumors, is complex and difficult to characterize. For example, immune gene expression as detected in high-throughput experiments originates from the combination of tumor, fibroblasts, endothelial, and immune cells in the microenvironment. Developing strategies to capture and quantify this immune component would facilitate the characterization of the poorly understood roles immunity plays in cancer progression. Currently, the approaches applied to profile the immune component of tumor fall short in achieving this goal. Analysis of the immune component of a tumor currently relies on incomplete identification of immune factors and their associated protein networks, primarily using manual

approaches. In **Paper II,** an immunological relevance score for all human genes, was developed using a novel strategy that combines literature mining and information theory. Using this score, an immunological grade can be assigned to gene expression profiles in a sample, and thereby quantify the immunological component of tumors.

Measures were taken to benchmark this score against existing manually curated immune resources, as well as against results from high-throughput studies. To further utilize immunological relevance for genes, the relevance score was charted against both protein networks and cancer information. This forms an expanded interactome landscape of tumor immunity. We applied this approach to expression profiles in melanomas, thus identifying and grading their immunological components, followed by identification of their associated protein networks.

The assignment of a ranked immunological relevance score to all human genes extends the content of existing immune gene resources and enriches our understanding of immune involvement in complex biological networks. The application of this approach to tumor immunity represents an automated systems strategy that quantifies the immunological component in complex disease.

Paper III

At the very core of complex protein networks in the cell are the biochemical mechanisms of function. These are enacted through protein complexes, or molecular machines. These units of functions are at the fundamental basis of all processes in the cell. To truly unravel the complexity of process such as Th cell differentiation, as approached in **Paper I**, it may be necessary to abstract the unit of function in the networks to that of the groups of protein that assemble as one entity to enact that function. This information however is incomplete, and although efforts such as those

in **Paper II** may help in acquiring more complete an accurate knowledge of the mechanisms, there is an abundance of ground yet to be achieved in understanding how these structure operate at the cellular level. It would therefore be highly informative to analyze cellular proteomes as networks of these interacting molecular machines. This was the first human study to treat true manually curated protein complexes, not their individual protein members, as the functional unit in cellular protein networks, and to predict interactions between them. To that end, we utilized expertly curated protein complexes and experimentally validated protein networks, and designed a statistical null model that randomized the membership of the protein complexes, but preserved their degrees in the protein networks. This statistical approach successfully identified the pairs of complexes in both human and yeast, where the number of corresponding protein interactions between them is due to an actual physical interaction between the complexes. An evaluation against a set of expertly curated yeast complex-complex interactions revealed that approximately 50% of these relationships could be predicted in this manner. A network analysis of high scoring complex-complex networks revealed a biologically sensible organization of the cell into functional networks of complex-complex interactions. Such high order analyses of cellular proteomes can lead to improved understanding of little understood cellular processes, and guide the discovery of novel relationships between molecular machines.

Discussion

Methodological & Biological perspectives

Validity of the Boolean networks

The Boolean network approach applied in **Paper I**[246] was more comprehensive than those attempted previously on T-cell differentiation [248]. The Th1 and Th2 cells have a phenotype defined by the release of individual signature cytokines, *i.e.* IFN-γ and IL-4, respectively. Previously it was thought that these cell types exist in a dichotomy whereby they co-regulate each other to coordinate the immune response[239,240]. However, recent evidence suggests that these two Th cell subtypes do not behave in a dichotomy of counter regulation. A novel algorithm was applied that allows the Boolean analysis of large networks[269] to model this phenotype outcome. It could be argued this new algorithm was not taken full advantage of in our approach. Several formal methods can be used in order to analyze steady states of a Boolean network to find all attractors. We tested both Binary Decision Diagrams (BDD)[271] and the approach of Boolean Satisfiability (SAT)[269] which allows the modeling of much larger networks. Although we expand the existing model[248] of Th cell differentiation from 14 to 51 genes, alternative methods, such as BDD, may have worked just as well on a network of this size. However, BDD and other decision diagram-based algorithms have limited capacity due to the excessive memory requirements for large networks. The reason for the limitation to 51 genes from the total 403 genes identified from literature mining the 20 million articles in Medline, was based on careful manual curation. Only well-defined interactions for Th1/Th2 cell differentiation were included. The relevance of this curation is substantiated to a certain extent by the analyses of gene expression data from patients with immunological disease. However, the limitation of a manual curation process in assigning the Boolean model

construction, by building the positive (activation) versus negative (inhibition) interactions statements, may have reflected the bias of the current knowledge in the field toward the Th1/Th2 dichotomy. Therefore, there is scope for expansion of this study to a more complete Boolean model constructed from all known genes associated to Th cell differentiation, extracted from the literature, where the directionality of the positive and negative statements are incorporated. The painstaking task assigning these rules manually may be avoided during the development of more intelligent literature mining algorithms that can capture activation/inhibition relationship automatically and accurately from unstructured text. Efforts in **Paper II** to identify and quantify immune phenotypes from complex tissue could also be applied in order to expand the limitations of existing Boolean models to unravel the complexity of this process in Th cells.

Synchronous vs asynchronous Boolean updating

The method applied to the Th network was designed to identify the steady states of Boolean networks under synchronous updating. This is an update scheme that assumes equal timescales for all interactions in the network. Such an assumption seems unlikely in a real living cell, as there are inter- and intracellular interactions as well as enzyme reactions and transcription, dynamically occurring at different time points. It has been well-established[271,272] that under asynchronous updating, the same initial state may lead to different steady states or attractors. This asynchronous updating could lead to different and more biologically accurate results. However, the data on Th cell differentiation that is available is only qualitative. Due to this limitation, and the example set by previous attempts[248], synchronous updating was chosen. However, the modeling of Th cell differentiation using asynchronous updating is very important for future research directions. This could be performed

based on biological input data, such as time-series gene expression microarray studies of Th cells polarized into the various lineages.

Plasticity of the Th cell lineage

This fourth attractor identified in the Boolean network model in **Paper I** did not agree with a counter-regulatory role of Th cell differentiation. This finding, however, does comply with the most recent empirical evidence on the "plasticity" and heterogeneity of Th cell types and differentiation[238,241,273]. An increasing number of Th cell subsets have been discovered, for example Th3[274], Th9[275], Th17[276,277] and Th22 [278]. An important conceptual problem that comes with network modeling of this newly discovered range of Th cells based on one cytokine output is that each subset expresses up to thousands of different proteins. There is a very large portion of these proteins, which overlap any two Th subsets. Each Th subset may have substantial phenotype similarities apart from that exerted by its signature cytokine. Thus, sub-classification based on one cytokine/subset may only reflect a fraction of the functionality of each subset.

Th cells are not only heterogeneous, but also plastic in their lineage fates. Their signature cytokines are released in complex protein network patterns[242], so that one lineage may change to another and back to the original lineage [242]. This process, and the protein interaction networks that mediate it, is not completely understood, especially *in vivo*. The sub-classifications of Th cells are largely based on *in vitro* studies, under conditions that may or may not resemble those *in vivo*. For example, *in vitro* Th cell polarization is induced by a fraction of the proteins involved *in vivo*. Another problem is that *in vivo* cells other than Th cells may release Th polarizing proteins (epithelial cells, mast cells and eosinophils). Ideally, a protein network model should comprise representative interaction networks from all these cells. This is a

formidable challenge, given the complexity of each network and the limitations of currently available information. Improved methodologies for extraction of these protein networks from the literature and analysis of single cells in vivo are likely to contribute to addressing this challenge.

Master regulators of Th cell plasticity

As discussed above, the Th cell is complicated by Th plasticity and great heterogeneity. Traditionally, Th cell classification has been discrete, rather than continuous, and mainly based on individual cytokines or transcription factors. However, Th plasticity implies continuous and transitional states through the other Th lineages. These transitional states are currently poorly characterized. However, in the context of Th1 and Th2 cells studied in **Paper I,** GATA3 has been shown to induce a transition from Th1 to Th2 cells [279]. Since Th1 cells are associated with TBET and Th2 cells with GATA3, this implies that TBET and GATA3 may oscillate between Th1, Th2 and transitional forms. This is supported by the description of TBET at low levels in Th2 cells and GATA3 in Th1 cells [280,281]. Similarly, IRF4 and STAT1 are expressed in both Th1 and Th2 cells[282]. Thus, the group of cells referred to as ThX may represent a transitional state between Th1 and Th2 cells. However, there is major limitation to capture this plasticity in it entirety in **Paper I**. For one example, the exclusion of FOXP3 in the manual curation step is a limitation. This transcription factor is a key component for Treg cell differentiation[283]. Therefore, it could be argued for the presence of a too strong a bias in the network modeling approach on the Th1 and Th2 cell differentiation to capture the plasticity of the Th cell lineage. For a comprehensive understanding of how Th cells orchestrate immune responses in the tumor microenvironment, it would be appropriate to capture also the dynamics of Th17 and Tregs, given the evidence of their role in mediating breast cancer metastasis

to the brain[230]

The epigenetic mechanisms of Th cell fate

That which is not considered directly in the Boolean model of the Th regulatory network in **Paper I**, is the epigenetic mechanisms that govern the Th cell fate[284-286]. The fate and maintenance of a Th cell precursor is very much dependent on the genome organization of the cell at any given phenotype state[287]. Dynamic chromatin remodeling can result in increased accessibility of the cytokine genes responsible for differentiation to transcription factors. Furthermore, chromatin remodeling can induce the silencing of cytokine expression of the alternative phenotypes by restricting access to transcription factors, and through methylation of DNA[284,286]. Evidence of epigenetic factors in Th differentiation emerged as early as 2001, when assessment of Th cells removed from one set of polarizing conditions and placed in conditions to induce the alternative cytokine program, demonstrated that this plasticity to change was lost after three or four cell divisions[288]. The protein network model analyzed in **Paper I** falls short of analyzing these features, but future networks studies would necessitate the incorporation of this information in order to unravel the complexities of this process.

These epigenetic mechanisms, it could be argued, occur at the level of the molecular machine, *i.e.* the protein complex. Already inherent within the known network of protein interactions lies a great deal of complexity. Developing a formal representation of the complex network of interacting molecular agents in the cell is important to develop accurate simulations of the cell, and will lead to an improved understanding of its dynamics. This biochemical machinery of the cell is the most valid agent in network models. Also, when analyzing protein networks, it may be the

most informative. To accurately model these mechanisms, requires knowledge of how the protein complex machinery in nucleus compartments interact with each other, resulting in genome conformations that oscillate with signal transduction. We are some way away from achieving this in its entirety[39]. One step in progress towards this, however, is to have the ability simulate an accurate map of the protein complexes that in turn recruit and interact with other protein complexes to bring about this genome conformation. The approach developed in **Paper III** attempted to predict interactions between stable complexes and may possibly guide experimentation to identify these epigenetic relationships in the years to come.

Tumor tissue heterogeneity & complex protein networks

The current understanding of cancer biology places increased emphasis on understanding the Th differentiation networks studied in **Paper I**, and how they mediate tumor immune responses in the tumor microenvironment[224,240,289]. There is incomplete knowledge of the important molecular players, and how they interact in complex networks. However, increased consideration of microenvironment related question is driving immunotherapy forward[290]. For now, these mainly using agents that block immunosupression networks[291]. However, as knowledge of the pathway mechanisms in the tumor microenvironment increases, it should be possible to develop new cancer therapies that are safer and more efficacious with respect to the specific microenviroment of each cancer patients. Having a complete and accurate knowledge of the immune components in the microenvironment one can target specific patients having high levels of inflammatory molecules, cytokines, chemokines, tumor-infiltrating T cells (TILs), dendritic cells (DC) and/or macrophages. Developing a complete molecular understanding, and developing clinical applications, of immune-based cancer therapeutics require information on the

molecular networks that that will elicit a tumor-specific acute inflammatory response that induces tumor rejection. The tumor may be programmed through its complex molecular networks to express or release molecular signatures to activate IFN-γ/Th1 type tumor-killing immune responses,[270,292]. There is a need for computational strategies, like that of **Paper II**, to capture relevant networks of proteins to guide an understanding of the relevant cancer protein pathways and development of new treatments. By doing so, we can broaden our knowledge of the complex processes governing tumor immune responses.

Information theoretic scoring of immune signals

The immune environment of complex tissue dictates the stimuli received by Th and other immune cells in the peripheral lymph nodes or in thymus and their differentiation lineages[293]. Capturing and quantifying the immune information in heterogeneous tissue is a challenge and was addressed in **Paper II.** Phenotype information for the molecular players in complex cellular networks is latent in the 20 million-plus articles of the medical literature, and it is reasonable that this could be used as a resource to quantify the immune phenotypes in genes in complex tissue. Medline is a rich resource of semantic information. However, it is a noisy communication channel emitting convoluted phenotype information for thousands of genes and much of the gene information in Medline is convoluted among multiple overlapping phenotypes. The principles of information theory and Shannon's entropy were applied to deal with this challenge[294]. Shannon's entropy was seen as a sensible measure to score the phenotype information content for genes from the literature. It required a once-off manual effort to achieve the task of building a lexicon of expertly chosen terms relevant for immunity. This lexicon was then treated as an information coding system for immune relevant signals. Thus, the literature association between a

gene and an immune phenotype term was treated as the observance of a "symbol" communicating the immune signal for that gene.

Biases in the information scoring of immune phenotypes

Many biases potentially exist in adapting such an approach to quantify phenotypes. For example, the many association captured from the citation principle that was used[113] are possibly contrary to the relationship of the gene to phenotype being assumed in the indexed articles. It is difficult to solve this entirely and precisely using automated approaches. However, measures could have been taken to apply other principles of literature association than co-citation alone. Another bias that may affect the contextual information content of a gene is the citation popularity of the gene itself in all of Medline, *i. e.* its probability of co-citation amongst all structured vocabulary terms (association to all possible phenotypes) in the entire information space of Medline. This bias can be quantified to a certain extent and it was corrected using the Kullback-Leibler (KL) or "relative entropy"[295]. This correction created a more accurate measure of information content that was used as an immunological score for each gene.

Multiple phenotypes in the tumor microenvironment

The immune information score is a global measure encompassing 1921 immune terms from structured vocabularies or ontologies. It was not a score attributed to a specific phenotype or direction of the immune response. It was demonstrated to accurately rank and identify global immune network signatures, and suggest new genes that could be populated into immune databases. It was successfully applied to classify melanoma patient groups[227,296], whose immune score corresponds to key clinical features of cancer progression or survival; in addition to identify the immune related

protein networks in these tumor microenvironments. The global score has also been used to detect an elevated immune response, accompanied by associated protein networks, within a group of early-onset colorectal cancer patients compared to a late onset colorectal cancer group[297]. However, that which is not achieved by the approach developed in **Paper II** is a quantification of immune information with respect to specific immune responses, and/or the directionality of the immune response. The immune score does not in an "unsupervised" manner, or without manual interpretation, detect the directionality of the immune response toward the tumor. The approach can be developed further, in an improved version, to assess whether the immune network signatures that it captures in the high-throughput measurements of tumor tissue are correlating with Th1 type tumor-killing responses, or with Th2 type tumor-promoting responses. There is a range of machine learning methods[298] that could have been applied to detect these network signatures. There is a strong possibility that there is a lack of adequately annotated high-throughput experiments to successfully carry out such an endeavor. However, at the very least, the immune information score could be segmented into more resolute and discrete classes of immune phenotypes (NK, MI/Macrophage activity, etc). These then can be quantified in a similar scoring scheme, and related to their protein networks signatures to elucidate possible mechanisms behind these immune responses in tumors.

Community detection in complex-complex networks

Having identified high-confidence complex-complex interactions in the cell, it creates an opportunity to explore how such higher order relationships in protein networks relate to "communities" of interacting molecular machines. There were many possible options available to explore the communities of predicted complex-complex interactions. Community detection of groups of nodes in real networks has become a

central quest in network studies in the recent decade. [299-303]. These have proven to be effective, although most do not capture features of how network components behave toward each other in real networks or systems in nature. This is also true for molecular cellular networks. In contrast to these previous methods in community detection, which have entirely focused on the grouping of nodes, a community detection algorithm was used in this study that naturally incorporated overlaps between the interactions between the molecular machines. In this way, a sensible biological organization of the interactome into functional communities of cooperating protein complexes is revealed.

It was apparent that these interactions organized the complexity of cellular protein networks into sensible biological clusters. It would have been additional supportive evidence to this effect, if we had measured how these communities of complex-complex interactions were organized according to their cellular compartments. This was not done systematically using statistical measures of ontology enrichment, for example, mostly because the majority of the complexes in the source databases used were nuclear complexes, which would have biased such tests *a priori*.

Additional observations from the protein complexome

In the course of building the data framework to predict complex-complex interactions, a number of observations were made which were either relevant to the specific goal of **Paper III**, or interesting in themselves and possible topics for future studies. One of these was the great degree of similarity between complexes in humans. Similarity was quantified between two complexes in terms of the number of overlapping proteins. This was computed by the Jaccard index, and then complex pairs were clustered based on this similarity. There was a great degree of such clustering, with a large

amount of clusters in the human proteome. In yeast, however, clusters of similar complexes were small. Similarity in protein interaction partners in the protein network was also interesting. Proteins in the protein networks of the cell were considered "similar" if they had the same interaction partners (also quantified using the Jaccard index). The extent to which similar proteins exist and were found within the same protein complex was notable. It was generally found that proteins that had high degree of similarity in their interaction partners were enriched in the largest complexes. The possible genomic or evolutionary reasons for these phenomena were largely left unexplored in **Paper III**, and are open to questions for future studies.

Concluding Remarks

The statistical methodology applied to predicting complex-complex interactions, in some respects, addresses a real need to develop further bioinformatics and statistical tools that reliably capture the interconnected environment of the cell. It has become clear that the cell is replete with correlated patterns and complexities, and traditional tools fall short. Progress towards robust network approaches is limited by the incompleteness of the entire network maps of the cell, and knowledge of the biochemical mechanisms that underlie their function.

Future perspectives

Designer circuits for personal cancer immunotherapy

The future development in protein complex design and the emerging field of network biology, encompasses promising strategies to both understand the complex networks in the cell, and also to deliver targeted therapies to cancer patients. This may have seemed unrealistic and overly ambitious only some few years ago, and as we are still many years to achieve this, it still may be considered an ambitious vision. However, it will be possible eventually to engineer functional molecular machines cooperating in complex cellular networks in living systems.

The merging of the fields of nanotechnology, bioinformatics, systems biology and synthetic biology can bring about a revolution in targeted therapies in cancer. This can be explained by the following hypothetical scenarios: (1) When we understand, finally, the complete dynamics of gene regulation in eukaryotic cells, using the toolkits developed in systems biology and brought to our comprehension in computational biology. (2) Using this acquired knowledge of gene circuitry in cells, we engineer synthetic circuits that will engage in a certain gene expression programs leading to the building of cellular machines that will carry out therapeutic actions through interactions with other cellular systems or cellular machines. (3) We use technologies developed in nanotechnology to effectively deliver these designer circuits to the target cells and tissues of a cancer patient. This will bring about true personal targeted molecular cancer medicine.

Realistic possibility to harness complexity

Designer T cells encoding antigen bound molecular machines have recently achieved clinical effect. It has recently been reported how immunotherapy with these designer cells, derived from cells from the patient, can recognize and destroy malignant cells.[304] Synthetic biology is developing rapidly towards harnessing gene circuits of increasing complexity[305,306]. The desire, ultimately, to construct and control molecular machines, fuels one of the great endeavors of contemporary biochemistry. This will be a major achievement in the future global approaches of computational biochemistry. Protein biochemists are making rapid advances in understanding the 3D structure of molecular machines combining nanotechnology engineering principles to construct them[307,308]. Finally, nanotechnology is making rapid advances in drug delivery and imitation of cellular systems. As progress increases in understanding dynamic complex systems, mastery of structure, function and communication across the different protein machines will prove essential.

References

1 Alberts, B. Molecular biology of the cell: Reference edition: Volume 1. 1601 (2008).

2 Hooke, R. *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*. (Printed by J. Martyn and J. Allestry, 1665).

3 Dutrochet, H. & Baillière ((París)), J.-B. Recherches anatomiques et physiologiques sur la structure intime ... 233 (1824).

4 Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651-654, doi:10.1038/35036627 (2000).

5 Garrod, A. E. *Inborn errors of metabolism*. 2d edn, (H. Frowde and Hodder & Stoughton, 1923).

6 Horowitz, N. H. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci USA* **31**, 153-157 (1945).

7 Beadle, G. W. & Tatum, E. L. in *Proc Natl Acad Sci USA* Vol. 27 499-506 (1941).

8 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

9 Vogel, F. A Preliminary Estimate of the Number of Human Genes. *Nature* **201**, 847 (1964).

10 Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36, doi:gkn721 [pii]
10.1093/nar/gkn721 (2009).

11 Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36**, D1009-1014, doi:gkm965 [pii]
10.1093/nar/gkm965 (2008).

12 Kandpal, R., Saviola, B. & Felton, J. The era of 'omics unlimited. *Biotechniques* **46**, 351-352, 354-355, doi:000113137 [pii]
10.2144/000113137 (2009).

13 Barabasi, A.-L. Scale-Free Networks: A Decade and Beyond. *Science* **325**, 412-413, doi:10.1126/science.1173299 (2009).

14 Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**, 56-68, doi:10.1038/nrg2918 (2011).

15 Pawson, T. & Linding, R. Network medicine. *FEBS Lett* **582**, 1266-1270, doi:10.1016/j.febslet.2008.02.011 (2008).

16 Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* **18**, 326-332, doi:10.1038/73786 (2000).

17 Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**, 135, doi:10.1038/msb4100177 (2007).

18 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355-360, doi:10.1093/nar/gkp896 (2010).

19 Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **104**, 1777-1782, doi:10.1073/pnas.0610772104 (2007).
20 Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213, doi:10.1186/1471-2105-11-213 (2010).
21 Jamshidi, N., Edwards, J. S., Fahland, T., Church, G. M. & Palsson, B. O. Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics (Oxford, England)* **17**, 286-287 (2001).
22 Edwards, J. S., Ramakrishna, R. & Palsson, B. O. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng* **77**, 27-36 (2002).
23 Gille, C. *et al.* HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* **6**, 411, doi:10.1038/msb.2010.62 (2010).
24 Zhu, Q. *et al.* Chemical basis of metabolic network organization. *PLoS Computational Biology* **7**, e1002214, doi:10.1371/journal.pcbi.1002214 (2011).
25 Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M. & Snyder, M. Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell* **143**, 639-650, doi:10.1016/j.cell.2010.09.048 (2010).
26 Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911-920, doi:10.1038/nature09645 (2010).
27 Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* **138**, 314-327, doi:10.1016/j.cell.2009.04.058 (2009).
28 Vermeirssen, V. *et al.* Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. *Genome Res* **17**, 1061-1071, doi:10.1101/gr.6148107 (2007).
29 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680, doi:10.1038/nrg2641 (2009).
30 Gottardo, R. Modeling and analysis of ChIP-chip experiments. *Methods Mol Biol* **567**, 133-143, doi:10.1007/978-1-60327-414-2_9 (2009).
31 Jong, H. d. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *http://dx.doi.org/10.1089/10665270252833208*.
32 Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* **9**, 770-780, doi:10.1038/nrm2503 (2008).
33 Kauffman, S. A. in *J Theor Biol* Vol. 22 437-467 (1969).
34 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840, doi:nature09267 [pii]
10.1038/nature09267 (2010).
35 Cekaite, L., Clancy, T. & Sioud, M. Increased miR-21 expression during human monocyte differentiation into DCs. *Front Biosci (Elite Ed)* **2**, 818-828 (2010).
36 Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141, doi:S0092-8674(10)00245-X [pii]

10.1016/j.cell.2010.03.009 (2010).
37 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356 (1961).
38 Kirschner, M. W. *et al.* Fifty years after Jacob and Monod: what are the unanswered questions in molecular biology? *Molecular Cell* **42**, 403-404 (2011).
39 Rajapakse, I. & Groudine, M. On emerging nuclear order. *J Cell Biol* **192**, 711-721, doi:jcb.201010129 [pii]
10.1083/jcb.201010129 (2011).
40 D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* **694**, 49-61, doi:10.1007/978-1-60761-977-2_4 (2011).
41 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691-697, doi:gkq1018 [pii]
10.1093/nar/gkq1018 (2011).
42 Soh, D., Dong, D., Guo, Y. & Wong, L. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* **11**, 449, doi:1471-2105-11-449 [pii]
10.1186/1471-2105-11-449 (2010).
43 Friedman, A. & Perrimon, N. Genetic screening for signal transduction in the era of network biology. *Cell* **128**, 225-231, doi:10.1016/j.cell.2007.01.007 (2007).
44 Adler, E. M. 2010: signaling breakthroughs of the year. *Sci Signal* **4**, eg1, doi:10.1126/scisignal.2001770 (2011).
45 Cusick, M. E. Interactome: gateway into systems biology. *Human Molecular Genetics* **14**, R171-R181, doi:10.1093/hmg/ddi335 (2005).
46 Bader, S., Kühner, S. & Gavin, A.-C. Interaction networks for systems biology. *FEBS Lett* **582**, 1220-1224, doi:10.1016/j.febslet.2008.02.015 (2008).
47 Borisov, N. *et al.* Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. *Mol Syst Biol* **5**, 256, doi:10.1038/msb.2009.19 (2009).
48 Firestein, R. *et al.* CDK8 is a colorectal cancer oncogene that regulates β-catenin activity. *Nature* **455**, 547-551, doi:10.1038/nature07179 (2008).
49 Fraser, I. D. C. & Germain, R. N. Navigating the network: signaling cross-talk in hematopoietic cells. *Nat Immunol* **10**, 327-331, doi:10.1038/ni.1711 (2009).
50 Bernards, R. Cancer: Entangled pathways. *Nature* **455**, 479-480, doi:10.1038/455479a (2008).
51 Heidorn, S. J. *et al.* Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. *Cell* **140**, 209-221, doi:10.1016/j.cell.2009.12.040 (2010).
52 Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* **39**, D261-267, doi:gkq1104 [pii]
10.1093/nar/gkq1104 (2011).
53 Gnad, F., Gunawardena, J. & Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* **39**, D253-260, doi:gkq1159 [pii]
10.1093/nar/gkq1159 (2011).

54 Chernorudskiy, A. L. *et al.* UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* **8**, 126, doi:1471-2105-8-126 [pii]
10.1186/1471-2105-8-126 (2007).
55 Linding, R. *et al.* NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* **36**, D695-699, doi:10.1093/nar/gkm902 (2008).
56 Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415-1426, doi:10.1016/j.cell.2007.05.052 (2007).
57 Venancio, T. M., Balaji, S., Iyer, L. M. & Aravind, L. Reconstructing the ubiquitin network: cross-talk with other systems and identification of novel functions. *Genome Biol* **10**, R33, doi:10.1186/gb-2009-10-3-r33 (2009).
58 Dephoure, N. *et al.* A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A* **105**, 10762-10767, doi:0805139105 [pii]
10.1073/pnas.0805139105 (2008).
59 Xu, G., Paige, J. S. & Jaffrey, S. R. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat Biotechnol* **28**, 868-873, doi:nbt.1654 [pii]
10.1038/nbt.1654 (2010).
60 Arntzen, M. O. & Thiede, B. ApoptoProteomics: An integrated database for analysis of proteomics data obtained from apoptotic cells. *Mol Cell Proteomics*, doi:M111.010447 [pii]
10.1074/mcp.M111.010447 (2011).
61 Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat Meth* **6**, 83-90, doi:10.1038/nmeth.1280 (2009).
62 Dreze, M. *et al.* High-quality binary interactome mapping. *Meth Enzymol* **470**, 281-315, doi:10.1016/S0076-6879(10)70012-4 (2010).
63 Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968, doi:10.1016/j.cell.2005.08.029 (2005).
64 Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178, doi:10.1038/nature04209 (2005).
65 Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-110, doi:10.1126/science.1158684 (2008).
66 Collins, S. R. *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-450, doi:10.1074/mcp.M600381-MCP200 (2007).
67 Fiedler, D. *et al.* Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-963, doi:10.1016/j.cell.2008.12.039 (2009).
68 Lemmens, I., Lievens, S. & Tavernier, J. Strategies towards high-quality binary protein interactome maps. *J Proteomics* **73**, 1415-1420, doi:10.1016/j.jprot.2010.02.001 (2010).
69 Futschik, M. E., Chaurasia, G. & Herzel, H. Comparison of human protein-protein interaction maps. *Bioinformatics* **23**, 605-611, doi:btl683 [pii]
10.1093/bioinformatics/btl683 (2007).

70 Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120, doi:gb-2006-7-11-120 [pii]
10.1186/gb-2006-7-11-120 (2006).
71 Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-834, doi:10.1093/bioinformatics/bti115 (2005).
72 Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**, D532-539, doi:10.1093/nar/gkp983 (2010).
73 Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-531, doi:10.1093/nar/gkp878 (2010).
74 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Research* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).
75 Goel, R., Muthusamy, B., Pandey, A. & Prasad, T. S. K. Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Mol Biotechnol* **48**, 87-95, doi:10.1007/s12033-010-9336-8 (2011).
76 Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-2371, doi:10.1101/gr.1680803 (2003).
77 Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704, doi:10.1093/nar/gkq1116 (2011).
78 Bader, G. D., Betel, D. & Hogue, C. W. V. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research* **31**, 248-250 (2003).
79 Bader, G. D. *et al.* BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* **29**, 242-245 (2001).
80 Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M. & Wodak, S. J. Interaction databases on the same page. *Nat Biotechnol* **29**, 391-393, doi:10.1038/nbt.1867 (2011).
81 Aranda, B. *et al.* PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* **8**, 528-529, doi:10.1038/nmeth.1637 (2011).
82 Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* **2010**, baq023, doi:10.1093/database/baq023 (2010).
83 Razick, S., Magklaras, G. & Donaldson, I. M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405, doi:10.1186/1471-2105-9-405 (2008).
84 Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biology* **7**, 120, doi:10.1186/gb-2006-7-11-120 (2006).
85 Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* **100**, 4372-4376, doi:10.1073/pnas.0735871100 (2003).
86 Wang, R.-S., Wang, Y., Wu, L.-Y., Zhang, X.-S. & Chen, L. Analysis on multi-domain cooperation for predicting protein-protein interactions. *BMC Bioinformatics* **8**, 391, doi:10.1186/1471-2105-8-391 (2007).
87 Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M. & Teichmann, S. A. Statistical analysis of domains in interacting protein pairs. *Bioinformatics* **21**, 993-1001, doi:10.1093/bioinformatics/bti086 (2005).

88 Hayashida, M., Ueda, N. & Akutsu, T. A simple method for inferring strengths of protein-protein interactions. *Genome Inform* **15**, 56-68 (2004).
89 Gomez, S. M., Lo, S. H. & Rzhetsky, A. Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* **159**, 1291-1298 (2001).
90 Han, D.-S., Kim, H.-S., Jang, W.-H., Lee, S.-D. & Suh, J.-K. PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Research* **32**, 6312-6320, doi:10.1093/nar/gkh972 (2004).
91 Itzhaki, Z., Akiva, E. & Margalit, H. Preferential use of protein domain pairs as interaction mediators: order and transitivity. *Bioinformatics* **26**, 2564-2570, doi:10.1093/bioinformatics/btq495 (2010).
92 Huang, C. *et al.* Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Trans Comput Biol Bioinform* **4**, 78-87, doi:10.1109/TCBB.2007.1001 (2007).
93 Guimarães, K. S., Jothi, R., Zotenko, E. & Przytycka, T. M. Predicting domain-domain interactions using a parsimony approach. *Genome Biol* **7**, R104, doi:10.1186/gb-2006-7-11-r104 (2006).
94 Guimarães, K. S. & Przytycka, T. M. Interrogating domain-domain interactions with parsimony based approaches. *BMC Bioinformatics* **9**, 171, doi:10.1186/1471-2105-9-171 (2008).
95 Pawson, T., Raina, M. & Nash, P. Interaction domains: from simple binding events to complex cellular behavior. *FEBS Lett* **513**, 2-10 (2002).
96 Lee, H., Deng, M., Sun, F. & Chen, T. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics* **7**, 269, doi:10.1186/1471-2105-7-269 (2006).
97 Kanaan, S. P., Huang, C., Wuchty, S., Chen, D. Z. & Izaguirre, J. A. Inferring protein-protein interactions from multiple protein domain combinations. *Methods Mol Biol* **541**, 43-59, doi:10.1007/978-1-59745-243-4_3 (2009).
98 Chen, L., Wu, L.-Y., Wang, Y. & Zhang, X.-S. Inferring protein interactions from experimental data by association probabilistic method. *Proteins* **62**, 833-837, doi:10.1002/prot.20783 (2006).
99 Riley, R., Lee, C., Sabatti, C. & Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol* **6**, R89, doi:10.1186/gb-2005-6-10-r89 (2005).
100 Deng, M., Mehta, S., Sun, F. & Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Research* **12**, 1540-1548, doi:10.1101/gr.153002 (2002).
101 Singhal, M. & Resat, H. A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics* **8**, 199, doi:10.1186/1471-2105-8-199 (2007).
102 Han, D., Kim, H.-S., Seo, J. & Jang, W. A domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Inform* **14**, 250-259 (2003).
103 Sprinzak, E. & Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol* **311**, 681-692, doi:10.1006/jmbi.2001.4920 (2001).
104 Jothi, R., Cherukuri, P. F., Tasneem, A. & Przytycka, T. M. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-

domain interactions mediating protein-protein interactions. *J Mol Biol* **362**, 861-875, doi:10.1016/j.jmb.2006.07.072 (2006).
105 Akiva, E., Itzhaki, Z. & Margalit, H. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc Natl Acad Sci USA* **105**, 13292-13297, doi:10.1073/pnas.0801207105 (2008).
106 Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453, doi:10.1126/science.1087361 (2003).
107 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-568, doi:10.1093/nar/gkq973 (2011).
108 Tsai, F. S. Text mining and visualisation of Protein-Protein Interactions. *Int J Comput Biol Drug Des* **4**, 239-244, doi:10.1504/IJCBDD.2011.041412 (2011).
109 Ananiadou, S., Kell, D. B. & Tsujii, J.-i. Text mining and its potential applications in systems biology. *Trends Biotechnol* **24**, 571-579, doi:10.1016/j.tibtech.2006.10.002 (2006).
110 Blagosklonny, M. V. & Pardee, A. B. Conceptual biology: unearthing the gems. *Nature* **416**, 373, doi:10.1038/416373a (2002).
111 Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7**, 119-129, doi:10.1038/nrg1768 (2006).
112 Stapley, B. J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput*, 529-540 (2000).
113 Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* **28**, 21-28, doi:10.1038/88213 (2001).
114 Donaldson, I. *et al.* PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 11 (2003).
115 Zhou, D., He, Y. & Kwoh, C. K. Extracting Protein-Protein Interactions from MEDLINE using the Hidden Vector State model. *Int J Bioinform Res Appl* **4**, 64-80 (2008).
116 Chowdhary, R., Zhang, J. & Liu, J. S. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics* **25**, 1536-1542, doi:10.1093/bioinformatics/btp245 (2009).
117 Daraselia, N. *et al.* Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604-611, doi:10.1093/bioinformatics/btg452 (2004).
118 Bandy, J., Milward, D. & McQuay, S. Mining protein-protein interactions from published literature using Linguamatics I2E. *Methods Mol Biol* **563**, 3-13, doi:10.1007/978-1-60761-175-2_1 (2009).
119 Müller, H.-M., Kenny, E. E. & Sternberg, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* **2**, e309, doi:10.1371/journal.pbio.0020309 (2004).

120 He, M., Wang, Y. & Li, W. PPI finder: a mining tool for human protein-protein interactions. *PLoS ONE* **4**, e4554, doi:10.1371/journal.pone.0004554 (2009).
121 Cusick, M. E. *et al.* Literature-curated protein interaction datasets. *Nat Meth* **6**, 39-46, doi:10.1038/nmeth.1284 (2009).
122 Kauffman, S. Homeostasis and differentiation in random genetic control networks. *Nature* **224**, 177-178 (1969).
123 Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* **22**, 437-467 (1969).
124 Kauffman, S. A. *Investigations*. (Oxford University Press, 2000).
125 Kaufman, M., Andris, F. & Leo, O. A logical analysis of T cell activation and anergy. *Proc Natl Acad Sci U S A* **96**, 3894-3899 (1999).
126 D'Ari, R. Thomas and D'Ari (1990) Biological feedback. *getcited.org* (1990).
127 Steggles, L. J., Banks, R., Shaw, O. & Wipat, A. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. *Bioinformatics* **23**, 336-343, doi:Doi 10.1093/Bioinformatics/Btl596 (2007).
128 Pogson, M., Smallwood, R., Qwarnstrom, E. & Holcombe, M. Formal agent-based modelling of intracellular chemical interactions. *BioSystems* **85**, 37-45, doi:Doi 10.1016/J.Biosystems.2006.02.004 (2006).
129 Barabási, A.-L. Scale-free networks: a decade and beyond. *Science* **325**, 412-413, doi:10.1126/science.1173299 (2009).
130 Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).
131 Barabasi, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).
132 Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41-42, doi:10.1038/35075138 (2001).
133 Goh, K.-I., Oh, E., Jeong, H., Kahng, B. & Kim, D. Classification of scale-free networks. *Proc Natl Acad Sci USA* **99**, 12583-12588, doi:10.1073/pnas.202301299 (2002).
134 Seebacher, J. & Gavin, A. C. SnapShot: Protein-protein interaction networks. *Cell* **144**, 1000, 1000 e1001, doi:S0092-8674(11)00177-2 [pii] 10.1016/j.cell.2011.02.025 (2011).
135 Breitkreutz, B.-J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biology* **4**, R22 (2003).
136 Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-2382, doi:10.1038/nprot.2007.324 (2007).
137 Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432, doi:10.1093/bioinformatics/btq675 (2011).
138 Hu, Z., Snitkin, E. S. & Delisi, C. VisANT: an integrative framework for networks in systems biology. *Briefings in Bioinformatics* **9**, 317-325, doi:10.1093/bib/bbn020 (2008).

139 Hu, Z. *et al.* VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Research* **37**, W115-W121, doi:10.1093/nar/gkp406 (2009).
140 Theocharidis, A., van Dongen, S., Enright, A. J. & Freeman, T. C. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc* **4**, 1535-1550, doi:10.1038/nprot.2009.177 (2009).
141 Brohee, S., Faust, K., Lima-Mendez, G., Vanderstocken, G. & van Helden, J. Network Analysis Tools: from biological networks to clusters and pathways. *Nat Protoc* **3**, 1616-1629, doi:nprot.2008.100 [pii]
10.1038/nprot.2008.100 (2008).
142 Nooy, W. d., Mrvar, A. & Batagelj, V. *Exploratory social network analysis with Pajek*. 2nd edn, (Cambridge University Press, 2011).
143 Schmidt, H. & Jirstrand, M. Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics* **22**, 514-515, doi:bti799 [pii]
10.1093/bioinformatics/bti799 (2006).
144 Ideker, T. & Sharan, R. Protein networks in disease. *Genome Res* **18**, 644-652, doi:10.1101/gr.071852.107 (2008).
145 Wang, X., Gulbahce, N. & Yu, H. Network-based methods for human disease gene prediction. *Brief Funct Genomics* **10**, 280-293, doi:10.1093/bfgp/elr024 (2011).
146 Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986-998, doi:10.1016/j.cell.2011.02.016 (2011).
147 Barabási, A.-L., Gulbahce, N. & Loscalzo, J. in *Nat Rev Genet* Vol. 12 56-68 (2011).
148 Chen, J., Aronow, B. J. & Jegga, A. G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* **10**, 73, doi:10.1186/1471-2105-10-73 (2009).
149 Oldenburg, R. A., Meijers-Heijboer, H., Cornelisse, C. J. & Devilee, P. Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol* **63**, 125-149, doi:10.1016/j.critrevonc.2006.12.004 (2007).
150 Bortoluzzi, S., Romualdi, C., Bisognin, A. & Danieli, G. A. Disease genes and intracellular protein networks. *Physiol Genomics* **15**, 223-227, doi:10.1152/physiolgenomics.00095.2003 (2003).
151 Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA* **105**, 4323-4328, doi:10.1073/pnas.0701722105 (2008).
152 Schuster-Böckler, B. & Bateman, A. Protein interactions in human genetic diseases. *Genome Biol* **9**, R9, doi:10.1186/gb-2008-9-1-r9 (2008).
153 Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90, doi:10.1038/nmeth.1280 (2009).
154 Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52, doi:10.1038/35011540 (1999).
155 Sakai, Y. *et al.* Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med* **3**, 86ra49, doi:10.1126/scitranslmed.3002166 (2011).

156 Soler-López, M., Zanzoni, A., Lluís, R., Stelzl, U. & Aloy, P. Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Res* **21**, 364-376, doi:10.1101/gr.114280.110 (2011).
157 Lage, K. *et al.* Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol Syst Biol* **6**, 381, doi:10.1038/msb.2010.36 (2010).
158 Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V. & Lankelma, J. Cancer: a Systems Biology disease. *BioSystems* **83**, 81-90, doi:10.1016/j.biosystems.2005.05.014 (2006).
159 Wang, E., Lenferink, A. & O'Connor-McCourt, M. Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol Life Sci* **64**, 1752-1762, doi:10.1007/s00018-007-7054-6 (2007).
160 Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205-4208, doi:10.1093/bioinformatics/bti688 (2005).
161 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).
162 Jonsson, P. F. & Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291-2297, doi:10.1093/bioinformatics/btl390 (2006).
163 Platzer, A., Perco, P., Lukas, A. & Mayer, B. Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* **8**, 224, doi:10.1186/1471-2105-8-224 (2007).
164 Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075, doi:10.1038/nature07423 (2008).
165 Torkamani, A. & Schork, N. J. Identification of rare cancer driver mutations by network reconstruction. *Genome Res* **19**, 1570-1578, doi:10.1101/gr.092833.109 (2009).
166 Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**, 1090-1098, doi:10.1038/ng1434 (2004).
167 Edelman, E. J., Guinney, J., Chi, J.-T., Febbo, P. G. & Mukherjee, S. Modeling cancer progression via pathway dependencies. *PLoS Computational Biology* **4**, e28, doi:10.1371/journal.pcbi.0040028 (2008).
168 Chang, J. T. *et al.* A Genomic Strategy to Elucidate Modules of Oncogenic Pathway Signaling Networks. *Molecular Cell* **34**, 104-114, doi:10.1016/j.molcel.2009.02.030 (2009).
169 Nibbe, R. K., Koyutürk, M. & Chance, M. R. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Computational Biology* **6**, e1000639, doi:10.1371/journal.pcbi.1000639 (2010).
170 Dutkowski, J. & Ideker, T. Protein networks as logic functions in development and cancer. *PLoS Computational Biology* **7**, e1002180, doi:10.1371/journal.pcbi.1002180 (2011).
171 Bezbradica, J. S. & Medzhitov, R. Integration of cytokine and heterologous receptor signaling pathways. *Nat Immunol* **10**, 333-339, doi:10.1038/ni.1713 (2009).

172 Lin, W.-W. & Karin, M. A cytokine-mediated link between innate immunity, inflammation, and cancer. *J. Clin. Invest.* **117**, 1175-1183, doi:10.1172/JCI31537 (2007).
173 Hu, X. & Ivashkiv, L. B. Cross-regulation of signaling pathways by interferon-gamma: implications for immune responses and autoimmune diseases. *Immunity* **31**, 539-550, doi:10.1016/j.immuni.2009.09.002 (2009).
174 Embrace the complexity. *Nat Immunol* **10**, 325, doi:10.1038/ni0409-325 (2009).
175 Davis, M. M. A prescription for human immunology. *Immunity* **29**, 835-838, doi:10.1016/j.immuni.2008.12.003 (2008).
176 von Herrath, M. G. & Nepom, G. T. Lost in translation: barriers to implementing clinical immunotherapeutics for autoimmunity. *J Exp Med* **202**, 1159-1162, doi:10.1084/jem.20051224 (2005).
177 Haining, W. N. & Wherry, E. J. Integrating genomic signatures for immunologic discovery. *Immunity* **32**, 152-161, doi:10.1016/j.immuni.2010.02.001 (2010).
178 Brusic, V., Zeleznikow, J. & Petrovsky, N. Molecular immunology databases and data repositories. *J Immunol Methods* **238**, 17-28 (2000).
179 Petrovsky, N. & Brusic, V. Computational immunology: The coming of age. *Immunol Cell Biol* **80**, 248-254, doi:10.1046/j.1440-1711.2002.01093.x (2002).
180 Pappalardo, F. *et al.* ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. *Briefings in Bioinformatics* **10**, 330-340, doi:10.1093/bib/bbp014 (2009).
181 Rapin, N., Lund, O. & Castiglione, F. Immune system simulation online. *Bioinformatics (Oxford, England)* **27**, 2013-2014, doi:10.1093/bioinformatics/btr335 (2011).
182 Rapin, N., Lund, O., Bernaschi, M. & Castiglione, F. Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system. *PLoS ONE* **5**, e9862 (2010).
183 Chaussabel, D. *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150-164, doi:10.1016/j.immuni.2008.05.012 (2008).
184 Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat Immunol* **10**, 116-125, doi:10.1038/ni.1688 (2009).
185 Kelley, J., de Bono, B. & Trowsdale, J. IRIS: a database surveying known human immune system genes. *Genomics* **85**, 503-511, doi:10.1016/j.ygeno.2005.01.009 (2005).
186 Lynn, D. J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol* **4**, 1-11, doi:10.1038/msb.2008.55 (2008).
187 Ortutay, C. & Vihinen, M. Immunome Knowledge Base (IKB): An integrated service for immunome research. *BMC Immunol* **10**, 3, doi:10.1186/1471-2172-10-3 (2009).

188 Clancy, T. *et al.* Immunological network signatures of cancer progression and survival. *BMC Med Genomics* **4**, 28, doi:10.1186/1755-8794-4-28 (2011).
189 Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273, doi:10.1371/journal.pgen.1001273 (2011).
190 Frankenstein, Z., Alon, U. & Cohen, I. R. The immune-body cytokine network defines a social architecture of cell interactions. *Biol Direct* **1**, 32, doi:10.1186/1745-6150-1-32 (2006).
191 Calvano, S. E. *et al.* A network-based analysis of systemic inflammation in humans. *Nature* **437**, 1032-1037, doi:10.1038/nature03985 (2005).
192 Hsueh, R. C. *et al.* Deciphering signaling outcomes from a system of complex networks. *Sci Signal* **2**, ra22, doi:10.1126/scisignal.2000054 (2009).
193 Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* **6**, 377, doi:10.1038/msb.2010.31 (2010).
194 Li, S., Wang, L., Berman, M., Kong, Y.-Y. & Dorf, M. E. Mapping a dynamic innate immunity protein interaction network regulating type I interferon production. *Immunity* **35**, 426-440, doi:10.1016/j.immuni.2011.06.014 (2011).
195 Thiery, J. P. & Sleeman, J. P. Complex networks orchestrate epithelial-mesenchymal transitions. *Nat Rev Mol Cell Biol* **7**, 131-142, doi:10.1038/nrm1835 (2006).
196 Korkaya, H., Liu, S. & Wicha, M. S. Regulation of Cancer Stem Cells by Cytokine Networks: Attacking Cancer's Inflammatory Roots. *Clin Cancer Res* **17**, 6125-6129, doi:10.1158/1078-0432.CCR-10-2743 (2011).
197 Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet* **6**, doi:10.1371/journal.pgen.1001113 (2010).
198 Nowotny, A. Beneficial effects of endotoxins. 581 (1983).
199 Virchow, R. L. in *Die Krankhaften Geschwülste. Dreissig Vorlesungen gehalten während des Wintersemesters 1862-63. Vierte Vorlesung.* Vol. 1 Ch. 4, 65 (1863).
200 Coley, W. B. The treatment of malignant tumors by repeated inoculations of erysipelas: with a report of ten original cases. *Am. J. Med. Sci.* **105**, 487–511 (1893).
201 Wiemann, B. & Starnes, C. O. Coley's toxins, tumor necrosis factor and cancer research: a historical perspective. *Pharmacol Ther* **64**, 529-564 (1994).
202 Ehrlich, P. Ueber den jetzigen Stand der Karzinomforschung. *Ned. Tijdschr Geneeskd* **5(Part 1)**, 273-290 (1909).
203 Dunn, G. P., Old, L. J. & Schreiber, R. D. The Three Es of Cancer Immunoediting. *Annu. Rev. Immunol.* **22**, 329-360, doi:10.1146/annurev.immunol.22.012703.104803 (2004).
204 Burnet, F. M. The concept of immunological surveillance. *Prog Exp Tumor Res* **13**, 1-27 (1970).

205 BURNET, M. Cancer: a biological approach. III. Viruses associated with neoplastic conditions. IV. Practical applications. *Br Med J* **1**, 841-847 (1957).
206 BURNET, M. IMMUNOLOGICAL FACTORS IN THE PROCESS OF CARCINOGENESIS. *Br Med Bull* **20**, 154-158 (1964).
207 Salk, J. Immunological paradoxes: theoretical considerations in the rejection or retention of grafts, tumors, and normal tissue. *Ann N Y Acad Sci* **164**, 365-380 (1969).
208 Stutman, O. Tumor development after 3-methylcholanthrene in immunologically deficient athymic-nude mice. *Science* **183**, 534-536 (1974).
209 Rygaard, J. & Povlsen, C. O. The mouse mutant nude does not develop spontaneous tumours. An argument against immunological surveillance. *Acta Pathol Microbiol Scand B Microbiol Immunol* **82**, 99-106 (1974).
210 Maleckar, J. R. & Sherman, L. A. The composition of the T cell receptor repertoire in nude mice. *J Immunol* **138**, 3873-3876 (1987).
211 Atkins, M. B. *et al.* High-dose recombinant interleukin 2 therapy for patients with metastatic melanoma: analysis of 270 patients treated between 1985 and 1993. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **17**, 2105-2116 (1999).
212 Kaplan, D. H. *et al.* Demonstration of an interferon gamma-dependent tumor surveillance system in immunocompetent mice. *Proc Natl Acad Sci USA* **95**, 7556-7561 (1998).
213 Dighe, A. S., Richards, E., Old, L. J. & Schreiber, R. D. Enhanced in vivo growth and resistance to rejection of tumor cells expressing dominant negative IFN gamma receptors. *Immunity* **1**, 447-456 (1994).
214 Shankaran, V. *et al.* IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **410**, 1107-1111, doi:10.1038/35074122 (2001).
215 Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol* **3**, 991-998, doi:10.1038/ni1102-991 (2002).
216 Dunn, G. P., Koebel, C. M. & Schreiber, R. D. Interferons, immunity and cancer immunoediting. *Nat Rev Immunol* **6**, 836-848, doi:10.1038/nri1961 (2006).
217 Balkwill, F., Charles, K. A. & Mantovani, A. Smoldering and polarized inflammation in the initiation and promotion of malignant disease. *Cancer Cell* **7**, 211-217, doi:10.1016/j.ccr.2005.02.013 (2005).
218 Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. Cancer-related inflammation. *Nature* **454**, 436-444, doi:10.1038/nature07205 (2008).
219 Balkwill, F. & Mantovani, A. Inflammation and cancer: back to Virchow? *Lancet* **357**, 539-545, doi:10.1016/S0140-6736(00)04046-0 (2001).
220 De Visser, K. E., Eichten, A. & Coussens, L. M. Paradoxical roles of the immune system during cancer development. *Nat Rev Cancer* **6**, 24-37, doi:10.1038/nrc1782 (2006).
221 Lu, H., Ouyang, W. & Huang, C. Inflammation, a key event in cancer development. *Mol Cancer Res* **4**, 221-233, doi:10.1158/1541-7786.MCR-05-0261 (2006).

222 Hussain, S. P. & Harris, C. C. Inflammation and cancer: an ancient link with novel potentials. *Int J Cancer* **121**, 2373-2380, doi:10.1002/ijc.23173 (2007).
223 Disis, M. L. Immune Regulation of Cancer. *Journal of Clinical Oncology*, doi:10.1200/JCO.2009.27.2146 (2010).
224 Pagès, F. *et al.* Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene* **29**, 1093-1102, doi:10.1038/onc.2009.416 (2010).
225 Galon, J. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science* **313**, 1960-1964, doi:10.1126/science.1129139 (2006).
226 Dave, S. S. *et al.* Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N Engl J Med* **351**, 2159-2169, doi:10.1056/NEJMoa041869 (2004).
227 Bogunovic, D. *et al.* Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci USA* **106**, 20429-20434, doi:10.1073/pnas.0905139106 (2009).
228 Zhang, L. *et al.* Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N Engl J Med* **348**, 203-213, doi:10.1056/NEJMoa020177 (2003).
229 Zou, W. Immunosuppressive networks in the tumour environment and their therapeutic relevance. *Nat Rev Cancer* **5**, 263-274, doi:10.1038/nrc1586 (2005).
230 Pardoll, D. Metastasis-Promoting Immunity: When T Cells Turn to the Dark Side. *Cancer Cell* **16**, 81-82, doi:10.1016/j.ccr.2009.07.007 (2009).
231 Ruffolo, C. *et al.* Subclinical intestinal inflammation in patients with Crohn's disease following bowel resection: a smoldering fire. *J Gastrointest Surg* **14**, 24-31, doi:10.1007/s11605-009-1070-9 (2010).
232 Denardo, D. G., Johansson, M. & Coussens, L. M. Immune cells as mediators of solid tumor metastasis. *Cancer Metastasis Rev* **27**, 11-18, doi:10.1007/s10555-007-9100-0 (2008).
233 DeNardo, D. G. *et al.* CD4(+) T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages. *Cancer Cell* **16**, 91-102, doi:S1535-6108(09)00216-5 [pii]
10.1016/j.ccr.2009.06.018 (2009).
234 Porta, C. *et al.* Cellular and molecular pathways linking inflammation and cancer. *Immunobiology* **214**, 761-777, doi:10.1016/j.imbio.2009.06.014 (2009).
235 Kuilman, T. & Peeper, D. S. Senescence-messaging secretome: SMS-ing cellular stress. *Nat Rev Cancer* **9**, 81-94, doi:10.1038/nrc2560 (2009).
236 Kang, T.-W. *et al.* Senescence surveillance of pre-malignant hepatocytes limits liver cancer development. *Nature* **advance online publication**, doi:http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature10599.html#supplementary-information (2011).
237 Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* **112**, 1557-1569, doi:112/5/1557 [pii]
10.1182/blood-2008-05-078154 (2008).

238 Zhou, L., Chong, M. M. & Littman, D. R. Plasticity of CD4+ T cell lineage differentiation. *Immunity* **30**, 646-655, doi:S1074-7613(09)00198-8 [pii] 10.1016/j.immuni.2009.05.001 (2009).
239 Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A. & Coffman, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J Immunol* **136**, 2348-2357 (1986).
240 Kidd, P. Th1/Th2 balance: the hypothesis, its limitations, and implications for health and disease. *Altern Med Rev* **8**, 223-246 (2003).
241 Zhu, J. & Paul, W. E. Heterogeneity and plasticity of T helper cells. *Cell Res* **20**, 4-12, doi:cr2009138 [pii] 10.1038/cr.2009.138 (2010).
242 Murphy, K. M. & Stockinger, B. Effector T cell plasticity: flexibility in the face of changing circumstances. *Nat Immunol* **11**, 674-680, doi:ni.1899 [pii] 10.1038/ni.1899 (2010).
243 Reiner, S. L. Decision making during the conception and career of CD4+ T cells. *Nat Rev Immunol* **9**, 81-82, doi:10.1038/nri2490 (2009).
244 Kaiko, G. E., Horvat, J. C., Beagley, K. W. & Hansbro, P. M. Immunological decision-making: how does the immune system decide to mount a helper T-cell response? *Immunology* **123**, 326-338, doi:10.1111/j.1365-2567.2007.02719.x (2008).
245 Elements of a good decision. *Nat Immunol* **11**, 645, doi:ni0810-645 [pii] 10.1038/ni0810-645 (2010).
246 Pedicini, M. *et al.* Combining Network Modeling and Gene Expression Microarray Analysis to Explore the Dynamics of Th1 and Th2 Cell Regulation. *PLoS Computational Biology* **6**, e1001032 (2010).
247 Naldi, A., Carneiro, J., Chaouiya, C. & Thieffry, D. in *PLoS computational biology* Vol. 6 e1000912 (2010).
248 Mendoza, L. A network model for the control of the differentiation process in Th cells. *BioSystems* **84**, 101-114, doi:10.1016/j.biosystems.2005.10.004 (2006).
249 Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291-294 (1998).
250 Alberts, B. & Miake-Lye, R. Unscrambling the puzzle of biological machines: the importance of the details. *Cell* **68**, 415-420 (1992).
251 Brenner, S. Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 207-212, doi:10.1098/rstb.2009.0221 (2010).
252 Chiu, W., Baker, M. L. & Almo, S. C. Structural biology of cellular machines. *Trends Cell Biol* **16**, 144-150, doi:10.1016/j.tcb.2006.01.002 (2006).
253 Sali, A. NIH workshop on structural proteomics of biological complexes. *Structure* **11**, 1043-1047, doi:S0969212603001631 [pii] (2003).
254 Aloy, P. & Russell, R. B. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* **7**, 188-197, doi:10.1038/nrm1859 (2006).
255 Levchenko, A. Computational cell biology in the post-genomic era. *Mol Biol Rep* **28**, 83-89 (2001).

256 Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* **37**, 825-831, doi:10.1093/nar/gkn1005 (2009).
257 Wang, H. *et al.* A complex-based reconstruction of the Saccharomyces cerevisiae interactome. *Molecular & Cellular Proteomics* **8**, 1361-1381, doi:10.1074/mcp.M800490-MCP200 (2009).
258 Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Research*, doi:10.1093/nar/gkp914 (2009).
259 Ozbabacan, S. E., Engin, H. B., Gursoy, A. & Keskin, O. Transient protein-protein interactions. *Protein Eng Des Sel* **24**, 635-648, doi:gzr025 [pii]
10.1093/protein/gzr025 (2011).
260 Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G. & Orengo, C. Transient protein-protein interactions: structural, functional, and network properties. *Structure* **18**, 1233-1243, doi:S0969-2126(10)00303-5 [pii]
10.1016/j.str.2010.08.007 (2010).
261 Bhardwaj, N., Abyzov, A., Clarke, D., Shou, C. & Gerstein, M. B. Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci* **20**, 1745-1754, doi:10.1002/pro.710 (2011).
262 Devos, D. & Russell, R. B. A more complete, complexed and structured interactome. *Curr Opin Struct Biol* **17**, 370-377, doi:S0959-440X(07)00076-0 [pii]
10.1016/j.sbi.2007.05.011 (2007).
263 Aloy, P. & Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem Sci* **27**, 633-638, doi:S0968000402022041 [pii] (2002).
264 Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531-537, doi:10.1038/nature03239 (2005).
265 Gavin, A.-C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147, doi:10.1038/415141a (2002).
266 Consortium, A. I. M. Evidence for network evolution in an Arabidopsis interactome map. *Science* **333**, 601-607, doi:10.1126/science.1203877 (2011).
267 Malovannaya, A. *et al.* Analysis of the human endogenous coregulator complexome. *Cell* **145**, 787-799, doi:10.1016/j.cell.2011.05.006 (2011).
268 Keseler, I. M. *et al.* EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research* **39**, D583-590, doi:10.1093/nar/gkq1143 (2011).
269 Dubrova, E. & Teslenko, M. A SAT-based algorithm for finding attractors in synchronous Boolean networks. *IEEE/ACM Trans Comput Biol Bioinform* **8**, 1393-1399, doi:10.1109/TCBB.2010.20 (2011).
270 Matzinger, P. & Kamala, T. Tissue-based class control: the other side of tolerance. *Nat Rev Immunol* **11**, 221-230, doi:nri2940 [pii]
10.1038/nri2940 (2011).

271 Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* **24**, 1917-1925, doi:btn336 [pii]
10.1093/bioinformatics/btn336 (2008).
272 Chaves, M., Albert, R. & Sontag, E. D. Robustness and fragility of Boolean models for genetic regulatory networks. *Journal of Theoretical Biology* **235**, 431-449, doi:Doi 10.1016/J.Boolean.2005.01.023 (2005).
273 Naldi, A., Carneiro, J., Chaouiya, C. & Thieffry, D. Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Comput Biol* **6**, e1000912, doi:10.1371/journal.pcbi.1000912 (2010).
274 Sakaguchi, S. *et al.* Foxp3+ CD25+ CD4+ natural regulatory T cells in dominant self-tolerance and autoimmune disease. *Immunol Rev* **212**, 8-27, doi:IMR427 [pii]
10.1111/j.0105-2896.2006.00427.x (2006).
275 Ma, C. S., Tangye, S. G. & Deenick, E. K. Human Th9 cells: inflammatory cytokines modulate IL-9 production through the induction of IL-21. *Immunol Cell Biol* **88**, 621-623, doi:icb201073 [pii]
10.1038/icb.2010.73 (2010).
276 Stockinger, B., Veldhoen, M. & Martin, B. Th17 T cells: linking innate and adaptive immunity. *Semin Immunol* **19**, 353-361, doi:S1044-5323(07)00085-1 [pii]
10.1016/j.smim.2007.10.008 (2007).
277 Harrington, L. E. *et al.* Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nat Immunol* **6**, 1123-1132, doi:ni1254 [pii]
10.1038/ni1254 (2005).
278 Eyerich, S. *et al.* Th22 cells represent a distinct human T cell subset involved in epidermal immunity and remodeling. *J Clin Invest* **119**, 3573-3585, doi:40202 [pii]
10.1172/JCI40202 (2009).
279 Ouyang, W. *et al.* Stat6-independent GATA-3 autoactivation directs IL-4-independent Th2 development and commitment. *Immunity* **12**, 27-37, doi:S1074-7613(00)80156-9 [pii] (2000).
280 Zhu, J. *et al.* Conditional deletion of Gata3 shows its essential function in T(H)1-T(H)2 responses. *Nat Immunol* **5**, 1157-1165, doi:ni1128 [pii]
10.1038/ni1128 (2004).
281 Jenner, R. G. *et al.* The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. *Proc Natl Acad Sci U S A* **106**, 17876-17881, doi:0909357106 [pii]
10.1073/pnas.0909357106 (2009).
282 Lund, R., Aittokallio, T., Nevalainen, O. & Lahesmaa, R. Identification of novel genes regulated by IL-12, IL-4, or TGF-beta during the early polarization of CD4+ lymphocytes. *J Immunol* **171**, 5328-5336 (2003).
283 Huehn, J., Polansky, J. K. & Hamann, A. Epigenetic control of FOXP3 expression: the key to a stable regulatory T-cell lineage? *Nat Rev Immunol* **9**, 83-89, doi:10.1038/nri2474 (2009).
284 Sawalha, A. H. Epigenetics and T-cell immunity. *Autoimmunity* **41**, 245-252, doi:10.1080/08916930802024145 (2008).

285 Janson, P. C. J., Winerdal, M. E. & Winqvist, O. At the crossroads of T helper lineage commitment-Epigenetics points the way. *Biochim Biophys Acta* **1790**, 906-919, doi:10.1016/j.bbagen.2008.12.003 (2009).
286 Aune, T. M., Collins, P. L. & Chang, S. Epigenetics and T helper 1 differentiation. *Immunology* **126**, 299-305, doi:10.1111/j.1365-2567.2008.03026.x (2009).
287 Natoli, G. Maintaining Cell Identity through Global Control of Genomic Organization. *Immunity* **33**, 12-24, doi:Doi 10.1016/J.Immuni.2010.07.006 (2010).
288 Grogan, J. L. *et al.* Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets. *Immunity* **14**, 205-215, doi:S1074-7613(01)00103-0 [pii] (2001).
289 Hrubisko, M. *et al.* Immunity profile in breast cancer patients. *Bratisl Lek Listy* **111**, 20-26 (2010).
290 Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565-1570, doi:331/6024/1565 [pii]
10.1126/science.1203486 (2011).
291 Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* **363**, 711-723, doi:NEJMoa1003466 [pii]
10.1056/NEJMoa1003466 (2010).
292 Matzinger, P. Friendly and dangerous signals: is the tissue in control? *Nat Immunol* **8**, 11-13, doi:ni0107-11 [pii]
10.1038/ni0107-11 (2007).
293 Takahama, Y. Journey through the thymus: stromal guides for T-cell development and selection. *Nat Rev Immunol* **6**, 127-135, doi:nri1781 [pii]
10.1038/nri1781 (2006).
294 Shannon, C. E. A Mathematical Theory of Communication. *At&T Tech J* **27**, 379-423 (1948).
295 Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann Math Stat* **22**, 142-143 (1951).
296 Riker, A. I. *et al.* The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med Genomics* **1**, 13, doi:10.1186/1755-8794-1-13 (2008).
297 Agesen, T. H. *et al.* CLC and IFNAR1 are differentially expressed and a global immunity score is distinct between early- and late-onset colorectal cancer. *Genes and immunity*, doi:10.1038/gene.2011.43 (2011).
298 Vanneschi, L. *et al.* A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min* **4**, 12, doi:1756-0381-4-12 [pii]
10.1186/1756-0381-4-12 (2011).
299 Song, J. & Singh, M. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics* **25**, 3143-3150, doi:10.1093/bioinformatics/btp551 (2009).
300 Rhrissorrakrai, K. & Gunsalus, K. C. MINE: Module Identification in Networks. *BMC Bioinformatics* **12**, 192, doi:10.1186/1471-2105-12-192 (2011).

301 Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC Syst Biol* **1**, 24, doi:10.1186/1752-0509-1-24 (2007).
302 Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98-101, doi:10.1038/nature06830 (2008).
303 Chen, J. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* **22**, 2283-2290, doi:10.1093/bioinformatics/btl370 (2006).
304 Kalos, M. *et al.* T Cells with Chimeric Antigen Receptors Have Potent Antitumor Effects and Can Establish Memory in Patients with Advanced Leukemia. *Science Translational Medicine* **3**, doi:ARTN 95ra73
DOI 10.1126/scitranslmed.3002842 (2011).
305 Doktycz, M. J. & Simpson, M. L. Nano-enabled synthetic biology. *Molecular Systems Biology* **3**, 125, doi:10.1038/msb4100165 (2007).
306 Purnick, P. E. M. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* **10**, 410-422, doi:10.1038/nrm2698 (2009).
307 Browne, W. R. & Feringa, B. L. Making molecular machines work. *Nat Nanotechnol* **1**, 25-35, doi:10.1038/nnano.2006.45 (2006).
308 Kinbara, K. & Aida, T. Toward intelligent molecular machines: directed motions of biological and artificial molecules and assemblies. *Chem Rev* **105**, 1377-1400, doi:10.1021/cr030071r (2005).

I

OPEN ACCESS Freely available online

Combining Network Modeling and Gene Expression Microarray Analysis to Explore the Dynamics of Th1 and Th2 Cell Regulation

Marco Pedicini[1☯], Fredrik Barrenäs[2☯], Trevor Clancy[3☯], Filippo Castiglione[1], Eivind Hovig[3,4,5], Kartiek Kanduri[2], Daniele Santoni[6,1], Mikael Benson[2,7]*

1 Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche (CNR), Rome, Italy, **2** The Unit for Clinical Systems Biology, University of Gothenburg, Gothenburg, Sweden, **3** Department of Tumor Biology, Institute of Cancer Research, the Norwegian Radium Hospital, Oslo, Norway, **4** The Institute for Medical Informatics, Rikshospitalet, Oslo University Hospital, Oslo, Norway, **5** Department of Informatics, University of Oslo, Oslo, Norway, **6** Barcelona Institute for Research in Biomedicine (IRB), Barcelona Science Park, Barcelona, Spain, **7** Unit for Pediatric Allergology, Queen Silvia Children's Hospital, Gothenburg, Sweden

Abstract

Two T helper (Th) cell subsets, namely Th1 and Th2 cells, play an important role in inflammatory diseases. The two subsets are thought to counter-regulate each other, and alterations in their balance result in different diseases. This paradigm has been challenged by recent clinical and experimental data. Because of the large number of genes involved in regulating Th1 and Th2 cells, assessment of this paradigm by modeling or experiments is difficult. Novel algorithms based on formal methods now permit the analysis of large gene regulatory networks. By combining these algorithms with *in silico* knockouts and gene expression microarray data from human T cells, we examined if the results were compatible with a counter-regulatory role of Th1 and Th2 cells. We constructed a directed network model of genes regulating Th1 and Th2 cells through text mining and manual curation. We identified four attractors in the network, three of which included genes that corresponded to Th0, Th1 and Th2 cells. The fourth attractor contained a mixture of Th1 and Th2 genes. We found that neither *in silico* knockouts of the Th1 and Th2 attractor genes nor gene expression microarray data from patients with immunological disorders and healthy subjects supported a counter-regulatory role of Th1 and Th2 cells. By combining network modeling with transcriptomic data analysis and *in silico* knockouts, we have devised a practical way to help unravel complex regulatory network topology and to increase our understanding of how network actions may differ in health and disease.

Citation: Pedicini M, Barrenäs F, Clancy T, Castiglione F, Hovig E, et al. (2010) Combining Network Modeling and Gene Expression Microarray Analysis to Explore the Dynamics of Th1 and Th2 Cell Regulation. PLoS Comput Biol 6(12): e1001032. doi:10.1371/journal.pcbi.1001032

Editor: Richard Bonneau, New York University, United States of America

Received June 20, 2010; **Accepted** November 11, 2010; **Published** December 16, 2010

Funding: This work was funded by the European Commission (FP6-2005-NEST-PATH, No. 043241 - ComplexDis and FP7-2008, No 223367- MultiMod) and the Swedish Research Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mikael.benson@vgregion.se

☯ These authors contributed equally to this work.

Introduction

The immune system is composed of diverse cell populations, for example antigen-presenting cells, T and B lymphocytes as well as effector cells like eosinophils, mast cells and neutrophils. One type of T lymphocytes, called T helper (Th), has an important role in regulating this cellular network. Th cells can be further divided into Th1 and Th2 cells. Th1 and Th2 cells are thought to be mutually inhibitory and also to be involved in different diseases; Th1 cells are associated with autoimmune diseases, while Th2 cells are involved in allergies [1].

Although considered a simplification, the Th1/Th2 dichotomy is supported by a large body of experimental evidence [2]. However, studies of human diseases are more ambiguous in terms of the counter-regulatory roles of Th1 and Th2 cells. We and others have found that allergy, which is mainly thought to be a Th2 disease, can also be associated with Th1 responses [3,4]. One explanation could be that the Th1/Th2 paradigm is, to a large extent, based on studies of gene interactions in mice which may differ from those in humans, [5]. Another important aspect is that Th1 and Th2 cells interact in complex cellular networks that include several other T-cell subsets and cell types [5]. Ultimately, the balance between Th1 and Th2 cells is complicated to study experimentally, because it is the net result of altered interactions between multiple genes.

Gene expression microarray studies evidence that hundreds of genes are involved in the Th1/Th2 cell differentiation [6]. We and others have found that complex gene expression changes in diseases can be addressed by arranging the genes in networks [7–9]. These networks give an overview of the genes that are involved, as well as their interactions, but not the dynamics of network changes that result in phenotypic alterations like, for example, Th1 and Th2 cell differentiation. Recent studies of the dynamics of Th1 and Th2 cell differentiation using *in silico* modeling have to some extent supported a counter-regulatory role of Th1 and Th2 cells [10,11].

The gene networks used have been based on a relatively small, though relevant, number of genes and interactions. In the present work we applied an algorithm previously developed to analyze large gene regulatory networks to perform *in silico* studies based on a more comprehensive gene network model, which included a larger number of genes [12,13].

Author Summary

Different T helper (Th) cell subsets have an important role in regulating the immune response in inflammatory diseases. Th1 and Th2 cells are thought to counter-regulate each other, and alterations in their balance result in different diseases.This paradigm has been challenged by recent clinical and experimental data. Because of the large number of genes involved in regulating Th1 and Th2 cells, assessment of this paradigm by experiments or modelling is difficult. In this study, we combined novel algorithms for network analysis, *in silico* knockouts, and gene expression microarrays to examine if Th1 and Th2 cells had counter-regulatory roles. We constructed a directed network model of genes that regulated Th1 and Th2 cells through text mining and manual curation. We identified four cycles in the gene expression dynamics, three of which expressed genes that corresponded to Th0 (Th1/Th2 precursor), Th1 and Th2 cells. The fourth cycle contained the expression of a mixture of Th1 and Th2 genes. We found that neither *in silico* knockouts of the Th1 and Th2 attractor genes nor gene expression microarray data from patients and healthy subjects supported a counter-regulatory role of Th1 and Th2 cells.

The network was constructed by combining text mining from Medline (www.pubmed.com) based on seed genes and protein interaction data, with manual annotation. The aim of our study was to examine if the so-constructed network model was compatible with a counter-regulatory role of Th1 and Th2 cells from healthy humans as well as patients with different inflammatory diseases.

To achieve this we studied the effects of *in silico* knockouts on the model dynamics [14], together with analyses of gene expression microarray studies of T-cells from healthy controls and patients with different inflammatory diseases.

Results

Definition of a network model of Th1 and Th2 differentiation

We defined a gene regulatory network (GRN) model of the genes involved in Th1 and Th2 cell differentiation based on manual annotation and automated data mining of Medline abstracts. To ease inspection, this gene regulatory network was organized into four layers according to the sub-cellular localization of the genes (see Figure 1). Another reason for this exercise was to enable the network for usage in agent-based models, as in [15].

The extracellular layer included cytokines (IL-7, TNFSF4, IFN-γ, IL-12 and IL-18), the antigens, as well as two membrane-receptors expressed on antigen-presenting cells, namely CD80 and CD86. The membrane layer consisted of the T-cell receptor and cytokine receptors. The intracellular layer included signaling molecules as well as transcription factors. Finally, an extra-cellular layer consisted of autocrine cytokines (IL-4 and IFN-γ) and paracrine cytokines (IL-5 and IL-13).

Characterization of the attractors of the network

Gene regulatory networks (GRNs) can be represented as graphs where nodes represent genes that are either active or inactive. The state of the network is given by the combination of the activation state of all genes. Starting from a certain state, the upcoming configuration is computed by applying synchronously an updating rule. In general, since the number of possible states is *finite* (*i.e.*, q^N if N is the number of nodes, and q is the number of possible values of a node), and the dynamics is deterministic, then from a given initial state, the network can only evolve towards a limit cycle (*i.e.*, *attractor*) of length one or more (up to q^N-1).

In what follows, we go after Kauffman [15] by identifying the attractors of the network dynamics as differentiation phases of the cell, and the transformations between attractors as pathways of cell differentiation.

Using the algorithm in [12] (briefly discussed in the Materials and Methods section), we found that the GRN dynamics was characterized by four attractors, three of which corresponded to known Th subsets, namely Th0, Th1 and Th2. The remaining attractor, which we named ThX, contained both Th1 and Th2 genes (see Table 1).

The Th1 and Th2 attractors contained either Th1 or Th2 genes, an observation that was compatible with a counter-regulatory role of Th1 and Th2 cells. For example, the Th1 attractor contained the transcription factor TBET, which has been experimentally shown to induce the Th1 cytokine IFN-γ and inhibit the Th2 transcription factor GATA3, which, in turn, induces the Th2 cytokine IL-4. Conversely, GATA3 inhibits TBET and IFN-γ. Thus, the two transcription factors TBET and GATA3 play a key role in the counter-regulatory interaction between Th1 and Th2 [5]. However, the mixture of Th1 and Th2 genes in the ThX attractor did not agree with a counter-regulatory role between Th1 and Th2 cells. In particular, the state s_1 contained both IFN-γ and IL-4, while the state s_3 contained both TBET and GATA3 (Table 1). This suggested that the dynamics of the network had an important role in regulating the balance between Th1 and Th2 cells. This may correspond, *in vivo*, to the situation in which antigenic stimulation may be temporary or persisting, and result in different inflammatory responses [16].

In silico knockouts to model the dynamics of the network

We performed single gene *in silico* knockout experiments for all genes in the network, in order to monitor the behaviour of the attractors. In so doing, we distinguished two different settings, corresponding to a different activation modality of the input nodes (*i.e.*, those contained in the yellow box of Figure 1): *temporary*-stimulation and *persisting*-stimulation. In temporary stimulation we examined the effects of an impulse-like stimulation of the input genes, which means that those genes were considered active for a short and transient period of time, and were set off thereafter. In persisting stimulation instead, inputs were set on or off throughout the observation period. Persisting stimulation is equivalent to introducing self-loops on the input nodes of the GRN.

We computed the number of attractors for each single-gene knockout and for both activation modalities. We found that the median number of attractors per knocked out gene was 4 (range 3–9) for temporary stimulation whereas it was 604 (range 322–1664) for persisting stimulation, (Table 2).

Therefore, as a first observation we noted that, similarly to *in vivo* stimulation, the network dynamics differed greatly between temporary and persisting stimulation. Next, we proceeded to examine the counter regulatory dynamics of the Th1 and Th2 cells. This was done by testing the effects of *in silico* knockouts of intracellular genes in the Th0, Th1, Th2 and ThX attractors. We started by knocking out TBET and GATA3. If TBET and GATA3 were counter-regulatory, knocking out TBET would be expected to result in attractors mainly containing IL-4, but not IFN-γ, while the opposite would be expected after knocking out GATA3.

Firstly, we applied the temporary stimulation activation modality (Figure 2). Knocking out TBET resulted in attractors that contained both IL-4 and IFN-γ, either IFN-γ or IL-4, as well as attractors without IL-4 and IFN-γ. Knocking out IL-4 resulted in attractors that contained either IFN-γ or IL-4, as well as attractors without IL-4 and IFN-γ.

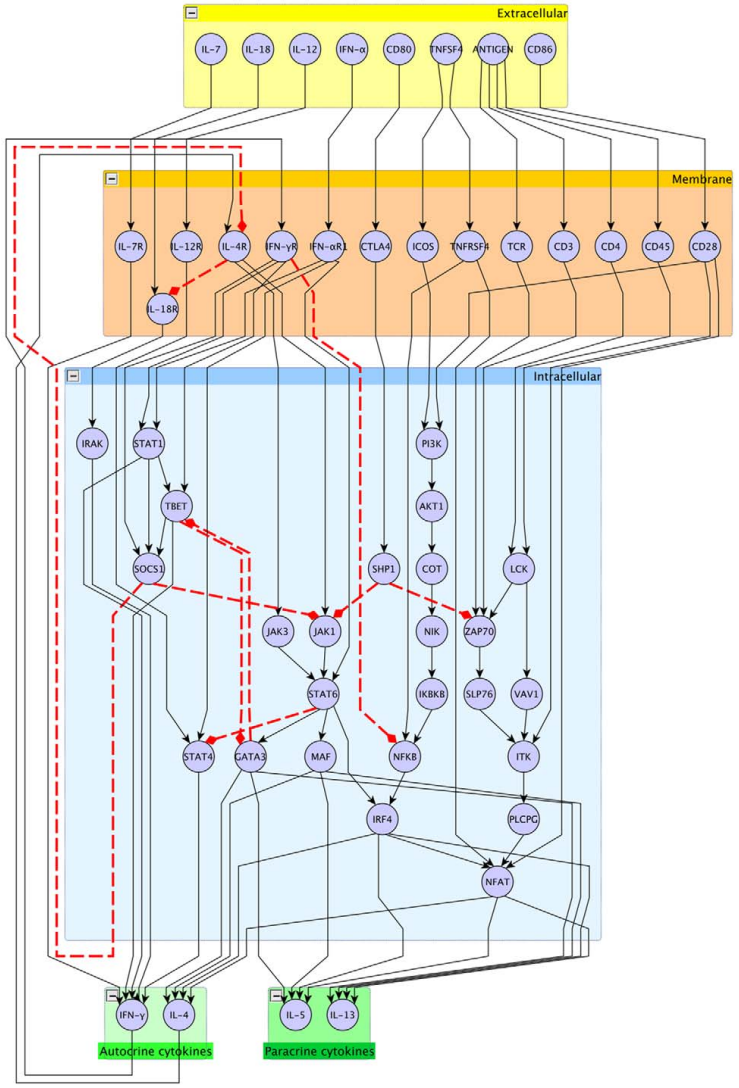


Figure 1. Systemic view of the gene regulatory network model including relevant genes or transcription factors for Th1 Th2 cell differentiation. Black edges depict positive regulation; red edges negative regulations.
doi:10.1371/journal.pcbi.1001032.g001

On the other hand, knocking out the same genes but applying the persisting stimulation activation modality mainly resulted in attractors containing both IL-4 and IFN-γ (Figure 3).

For both temporary and persisting stimulation, the knockout of other transcription factors that regulated Th1 and Th2 cells, namely IRF4, MAF, NFAT, STAT1 and STAT6, also resulted in attractors that contained IL-4 and IFN-γ, either alone or in combination. Thus, the balance between Th1 and Th2 cells was regulated by several transcription factors, and not only by TBET and GATA3.

To summarize, these findings were not compatible with a strictly counter-regulatory role of neither TBET nor GATA3 or any of the other transcription factors.

Analysis of relations between *in silico* and *in vitro* findings in human T-cells in health and disease

We proceeded to examine how the *in silico* findings related to *in vitro* studies of T-cells from healthy controls and patients with different T-cell related diseases. We downloaded several sets of gene expression microarray data from the public domain to test

Table 1. The attractors of the boolean network modeling Th1/Th2 differentiation.

Attractor	Active genes
Th0	None
Th1	IFN-γ, IFN-γR, SOCS1, STAT1, TBET
Th2	GATA3, IL-13, IL-4, IL-4R, IL-5, IRF4, JAK1, JAK3, MAF,
	NFAT and STAT6
ThX	s_1: IFN-γ, IL13, IL-4, IL-5, JAK3, NFAT and SOCS1
	s_2: IFN-γR, IL13, IL-4, IL-5 and STAT6
	s_3: GATA3, IL-4R, IRF4, MAF, SOCS1, STAT1 and TBET

ThX is the non-Th1-nor-Th2 attractor, consisting of a cycle composed by the three states s_1, s_2 and s_3.
doi:10.1371/journal.pcbi.1001032.t001

whether Th1 and Th2 genes were inversely correlated in T-cell related diseases.

If Th1 and Th2 cells are antagonists we would expect inverse relations between genes in the Th1 and Th2 attractors. If so, the expression levels of those genes would be negatively correlated. Instead of this, we found a highly significant positive correlation between the ratios of differentially expressed Th1-associated genes and Th2-associated genes (Pearson correlation coefficient $r = 0.799$, $p-\text{value} < 0.005$).

Thereafter, we analyzed the correlations between all gene pairs in the model that, based on the literature, were considered to inhibit each other. This analysis showed that all gene pairs were positively correlated but one (see Table 3).

This included the signature Th1 and Th2 genes TBET and GATA3, which showed the most significant positive correlation ($r = 0.81$, $p < 10^{-14}$) as well as IFN-γ and IL-4 ($r = 0.34$, $p < 0.01$).

Discussion

Because of the large number of proteins involved in Th cell differentiation, alterations in the balance between those proteins are not easily studied experimentally. Computational modeling provides an attractive alternative to study the dynamics of Th1 and Th2 cell regulation and has previously been employed for this purpose by us and others [10,11,17].

Such models have supported a counter-regulatory role of Th1 and Th2 cells, but were based on a relatively limited number of genes and did not include comparisons with biological data. In this report, we aimed to examine if Th1 and Th2 cells were counter-regulatory by combining modeling, *in silico* knockouts and gene expression microarray analyses of human T cells in health and disease. We constructed a network model of the proteins involved in Th cell differentiation by manual curation of proteins associated with Th1 and

Table 2. Number of attractors in knock-out networks.

	Temporary Stimulation			Sustained Stimulation		
knock-out gene	attractors	(static/dynamical)	max	attractors	(static/dynamical)	max
COT	4	(3/1)	3	1186	(898/288)	3
GATA3	3	(3/0)	1	322	(322/0)	1
IKBKB	4	(3/1)	3	594	(450/144)	3
IRAK	4	(3/1)	3	612	(452/160)	3
IRF4	9	(3/6)	5	604	(450/154)	5
ITK	4	(3/1)	3	1188	(900/288)	3
JAK1	4	(3/1)	3	594	(450/144)	3
JAK3	3	(3/0)	1	560	(432/128)	2
LCK	4	(3/1)	3	1187	(899/288)	3
MAF	4	(3/1)	3	594	(450/144)	3
NFAT	9	(3/6)	5	604	(452/152)	5
NFKB	4	(3/1)	3	594	(450/144)	3
NIK	4	(3/1)	3	1186	(898/288)	3
PI3K	4	(3/1)	3	1186	(898/288)	3
PLCPG	4	(3/1)	3	596	(452/144)	3
SHP1	4	(3/1)	3	594	(450/144)	3
SLP76	4	(3/1)	3	594	(450/144)	3
SOCS1	8	(5/3)	3	978	(594/384)	3
STAT1	7	(3/4)	6	1154	(482/672)	7
STAT4	4	(3/1)	3	612	(452/160)	3
STAT6	6	(3/3)	3	1664	(1088/576)	3
TBET	7	(3/4)	6	358	(322/36)	6
VAV1	4	(3/1)	3	595	(451/144)	3
ZAP70	4	(3/1)	3	1186	(898/288)	3

Number of attractors for the sustained and temporary stimulation; we give also the number of attractors which are of length one (*static equilibrium*) or of length greater than 1 (*dynamical equilibrium*); moreover, we indicate the maximal length of attractors.
doi:10.1371/journal.pcbi.1001032.t002

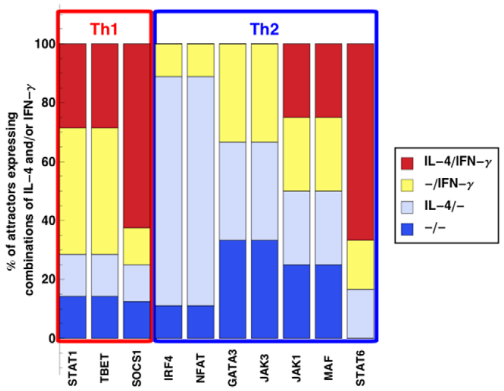


Figure 2. Number of attractors as the result of *in silico* knockout experiments, in the temporary stimulation activation modality. Stacked bars represent the percentage of attractors expressing combinations of IL-4 and/or IFN-γ.
doi:10.1371/journal.pcbi.1001032.g002

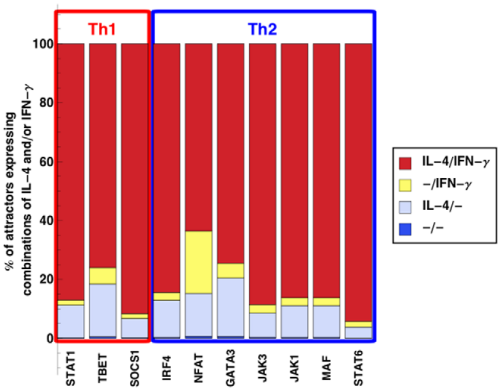


Figure 3. Number of attractors as the result of *in silico* knockout experiments, in the persisting stimulation activation modality. Stacked bars represent the percentage of attractors expressing combinations of IL-4 and/or IFN-γ.
doi:10.1371/journal.pcbi.1001032.g003

Th2 cells, and that had been identified as relevant through automated text mining of the medical literature. This resulted in a significantly more comprehensive model compared to previous versions.

Analysis of the dynamics of that model showed that it contained four attractors, two of which corresponded to the Th1 and Th2 subsets. These contained the Th1 and Th2 specific transcription factors TBET and GATA3, respectively. This was compatible with a counter-regulatory role of these attractors. However, the fourth attractor, which we named ThX, contained a mixture of Th1 and Th2 proteins, including TBET and GATA3. This did not agree with a counter-regulatory role of these transcription factors.

Furthermore, we extended our analysis by *in silico* knockout experiments of TBET and GATA3. We reasoned that if the two were counter-regulatory, then knocking out TBET would result in attractors mainly containing IL-4, while knocking out GATA3 would result in attractors mainly containing IFN-γ. Whereas this was true for GATA3, it was not the case for TBET.

In fact, knockout of either TBET or the other Th1 and Th2 attractor proteins mainly resulted in attractors containing both IFN-γ and IL-4. Afterthat, we examined the expression of Th1 and Th2 attractor genes in microarray studies of eleven T cell diseases, namely autoimmune, infectious and oncological diseases.

In most of these, the expression of Th1 and Th2 attractor genes increased concurrently, rather than in an opposing pattern. Moreover, we found that genes in the network model that were thought to inhibit each other based on experimental studies, were in fact positively correlated. This was particularly true for TBET and GATA3 which are thought to have a key role for the counter-regulation of Th1 and Th2 cells. It is of note that the interactions in the model were chosen based on experimentally validated functions and interactions in Th cells. In many cases those experiments were performed using polarizing cytokines and T cell receptor stimulants. This is likely to result in more homogenous Th cell responses than those seen *in vivo*. In the latter case Th cells are activated by antigen-presenting cells which process the antigens to peptides, subtle variants of which may have different effects on Th cells. In addition, different doses and timing of antigen exposure play an important role in the Th cell activation and differentiation process. The effects of timing was reflected by the results in our study; temporary and persistent stimulation had profound effects on the network dynamics of these processes.

Moreover, the activation involves a complex and variable mixture of proteins. Taken together, it is possible that this complexity may result in a mixture of Th1 and Th2 cells responses, rather than one of the two. The ThX attractor may correspond to such a mixed or transitional response. This is consistent with the increasing recognition that Th cell phenotypes are plastic rather than discrete [2]. This recognition resulted from experimental and clinical studies that show overlap between genes considered to be Th1 and Th2 genes [18,19].

Our analyses of gene expression microarray data from human T cells in health and disease lend further support to Th plasticity. From an *in vivo* perspective, this plasticity allows fine-tuned responses to a constant exposure of different antigens at different time points and doses.

It is also of note that *in vivo* Th1 and Th2 differentiation may be affected by many other T cell subsets, of which an increasing number have been recognized. Moreover, epithelial cells, mast cells and eosinophils release cytokines that affect the differentiation process. Ideally, simultaneous analysis of networks representing those cells and subsets would yield an understanding not only of Th1 and Th2 cells, but comprehensive models of the cellular networks that underlie immunological diseases. Improved meth-

Table 3. Results of Pearson's correlation test of inhibitory gene pairs.

Gene Pair	p-value	correlation r
GATA3 - TBET	$<10^{-14}$	0.81
SHP1 - JAK1	$<10^{-9}$	0.69
IL-4R - IL18R	$<10^{-5}$	0.56
SOCS1 - IL-4R	$<10^{-11}$	0.76
SOCS1 - JAK3	$=0.9341$	-0.01
STAT6 - STAT4	$<10^{-5}$	0.56
IFN-γ - IL-4	$=0.00702$	0.34

doi:10.1371/journal.pcbi.1001032.t003

odologies, such as single cell RNA sequencing may make such models feasible in the near future.

A limitation is that our model is that the underlying biological data is mainly qualitative. Thus, the model is based on synchronous updating and does not take into account quantitative or time-dependent changes. Others have shown that asynchronous updating may have different effects on attractors [20–22]. An interesting future research direction is to perform time series experiments of Th1 and Th2 cells using gene expression microarrays. Using such data it may be possible to improve our model both with regards to quantitative and time-dependent changes and also make predictions which can be validated with other biological methods, such as measuring Th1 and Th2 cytokines on the protein level.

In summary, our findings, both based on *in silico* modeling and analysis of T cells from human diseases agree with Th1 and Th2 cells having complex and possibly synergistic, rather than counter-regulatory roles.

Materials and Methods

Identification of genes for the network model of Th1/Th2 cell differentiation

We employed a step-wise procedure to define the set of relevant genes for the differentiation of Th cells into the Th1 and Th2 phenotypes.

Firstly, we identified two different sets of genes as a primary source: i) 17 genes from a previous network model [10]; ii) a set of 17 genes determined in a gene expression microarray study of polarized Th1 and Th2 cells by [6]. All these 34 genes were used as seed genes. Then we retrieved the first order neighbors of these seed genes and their connections in the BioGrid database (www.biogrid.org). Successively, the connection among the proteins of the first order neighbors were retrieved. Among all the genes retrieved thus far, we selected only those associated to the Gene Ontology term (www.geneontology.org) "T cell differentiation".

More specifically, the genes co-cited in the millions Medline abstracts together with this term were retrieved. This resulted in a set of 403 genes, that was further slimmed down and used to construct a manually annotated directed graph of gene interactions relevant for Th1 and Th2 cell differentiation. This was made by using the T-cell receptor pathway in the KEGG database as a template (www.genome.jp/kegg/pathway.html). Genes that were part of that pathway and had well-characterized and experimentally verified functions relevant for Th1 and Th2 cell differentiation were selected for the final network model. A detailed description of each interaction in the network, together with 126 supporting references is given in Text S1. It is also of note, that the network model was independent of the gene expression microarray experiments, which are described below (none of the published abstracts pertaining to those experiments contained co-cited genes that were included in the model).

Boolean networks as a model of Genetic Regulatory Networks

Given a GRN, the number of attractors of the network dynamics is, in general, not effectively computable since the number of states of the network grows exponentially with N. It is not even possible to effectively calculate the initial states of the network that will eventually fall in the basin of attraction of a specified limit cycle. When the number of genes is large, the explicit computation of the dynamics becomes impractical as the number of states the network can assume increases exponentially with the number of nodes. In the worst case the algorithm needs to store the complete description of the state transition graph and therefore the exhaustive study is feasible only when the number of nodes is small [10,23]. Just to give an idea, for a network with $N=40$ nodes, one needs about 6 Terabytes to store the state-transition graph of the network. In our case, with $N=51$, it would require about 7 Petabytes of storage.

In recent studies, formal methods such as *bounded model-checking technique* or *reduced order binary decision diagrams* have been employed in the study of attractors of Boolean and multivalued networks, see Dubrova *et al.*, Garg *et al.*, and Chaves *et al.* [12,21,24–26]. These formal methods have a potential to handle large networks. In particular we used Dubrova's algorithm based on a solver for the *satisfiability* problem (SAT) which from the logical structure of the network infers the possible attractors. In simple words, the network can be seen a Boolean circuit and its attractors can be computed by using methods and largely optimised algorithms coming from modeling of *Very Large Scale Integration* (VLSI) circuits.

What is special about formal methods approach is that it enables to find attractors of large networks. The idea behind the search algorithm is that, by using *symbolic* computation, it is possible to unfold the dependencies between nodes that are linked together and to compose the update function as a relation among the states (activation/inhibition) of the genes/nodes. Then the algorithm uses the SAT solver to determine the values of the states that *evaluate to true* the relation. This process is then repeated until all attractors are identified.

We specified the network as the set of rules $R_1, R_2, \ldots, R_N$, each one representing the activatory or inhibitory relation between genes. For example, if rule R stems for the activation of gene g, and is determined by the activators $x_1, x_2, \ldots, x_n$ and inhibitors $y_1, y_2, \ldots, y_m$ (activators and inhibitors are generically called regulators), then it can be written as $R: = x_1, \ldots, x_n, -y_1, \ldots, -y_m \rightarrow g$, where conventionally the subset of inhibitors are tagged with a minus sign.

Analogously to [10,12], the time is discrete and the activation states of the genes are changed simultaneously (*i.e.*, synchronous update). At each time step t, the value of the gene g is denoted by the same gene name $g(t)$. The successive value of gene $g(t+1)$ is

$$g(t+1) = \left(\bigvee_{i=1}^{n} x_i(t) \right) \wedge \neg \left(\bigvee_{j=1}^{m} y_j(t) \right) \qquad (1)$$

where $\vee, \wedge$ and $\neg$ denote the logical operators *and*, *or* and *not* respectively. The rule in Equation 1 states that for a gene to be activated, at least one activator and no inhibitors must be active [10 12, 27].

In our specific case we had a set of 43 rules involving 51 genes, that was the result of data mining and manual annotation. These are listed in Table 4. The network so specified was compiled in other formats, in particular GML (Graph Modeling Language), which is used in several applications specialized in graphical layout, and CNET which is the input form accepted by the algorithm to compute the attractors. Whereas the GML output was based simply on activation/inhibition network links, in the CNET format we had to specify the updating function for each node.

The last part of this work was the systematic characterization of the networks obtained by knocking out genes one at a time. As a consequence of these *in silico* knockout experiments we anticipated two results: a) to identify the set of genes which are pivotal to the Th1/Th2 differentiation; b) to spot subsets of co-expressed genes belonging to the attractors, since from analysis of microarray data we expected these genes to be correlated.

Changes in the set of the attractors were used to highlight relevant nodes. As a first approximation, differences in the mere number of attractors were considered; if a node did not affect the number of attractors, then from the point of view of the dynamics it was considered irrelevant.

Table 4. Specification rules for activation/inihibition links of the network in Figure 1.

IRF4, NFAT, MAF, GATA3	→	IL-13
IRF4, NFAT, MAF, GATA3	→	IL-5
IFN-γR, -GATA3, STAT1	→	TBET
IL-7R, TBET, STAT4,STAT1,IRAK	→	IFN-γ
STAT6	→	MAF
STAT6, -TBET	→	GATA3
IL-7	→	IL-7R
IL-18, -IL-4R	→	IL-18R
IL18R	→	IRAK
IFN-αR1, IFN-γR	→	STAT1
IFN-α	→	IFN-αR1
IFN-γ	→	IFN-γR
IL-4, -SOCS1	→	IL-4R
IRF4, NFAT, MAF, GATA3	→	IL-4
CD80	→	CTLA4
CTLA4	→	SHP1
CD45, CD4	→	LCK
TCR, CD3, -SHP1,LCK	→	ZAP70
ZAP70	→	SLP76
LCK	→	VAV1
CD28, VAV1, SLP76	→	ITK
ITK	→	PLCPG
ANTIGEN	→	CD4
ANTIGEN	→	TCR
ANTIGEN	→	CD3
ANTIGEN	→	CD45
TNFSF4	→	TNFRSF4
-IFN-γR, TNFRSF4, IKBKB	→	NFKB
STAT6, NFKB	→	IRF4
CD28, TNFRSF4, PLCPG, IRF4	→	NFAT
CD28, ICOS	→	PI3K
PI3K	→	AKT1
AKT1	→	COT
COT	→	NIK
NIK	→	IKBKB
CD86	→	CD28
IL-4R, -SHP1,-SOCS1	→	JAK1
IL-4R	→	JAK3
IFN-αR1, JAK1,JAK3	→	STAT6
IL12R, IFN-αR1, -STAT6	→	STAT4
IFN-γR, STAT1, TBET	→	SOCS1
IL-12	→	IL-12R
TNFSF4	→	ICOS

doi:10.1371/journal.pcbi.1001032.t004

Compilation and analysis of gene expression microarray data

To examine whether Th1 and Th2 gene activation patterns denoted two opposed pathways, gene expression data were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Datasets were selected based on the criteria that they i) measured mRNA expression from CD4+ cells from healthy controls or patients with T-cell related diseases (*e.g.*, SLE) and ii) that there were at least 5 samples per disease or per controls, (Table 5).

Table 5. Gene expression microarray datasets downloaded from the Gene Expression Omnibus repository.

GEO Accession Number	Disorder
GSE4588	Systemic Lupus Erythematosus (SLE), Rheumatoid Arthritis (RA)
GSE6740	HIV
GSE8835	B cell chronic lymphocytic leukemia (CLL)
GSE9927	Type I HIV (HIV-I)
GSE10586	Type 1 Diabetes (T1D)
GSE12079	Hypereosinophilic syndrome
GSE13732	Clinically Isolated Syndrome - Multiple Sclerosis
GSE14317	Adult T-cell leukemia/lymphoma (ATL)
GSE14924	Acute Myeloid Leukaemia (AML)
GSE17354	Adenosine deaminase (ADA) - Severe combined immunodeficiency (SCID) (Therapy treated)

doi:10.1371/journal.pcbi.1001032.t005

Differentially expressed genes between patients and controls in each disease were determined using the unpaired Student's t-test. Genes with a significance $p-\text{value}<0.05$ were considered differentially expressed.

In order to examine if the differentially expressed genes in the Th1 and Th2 attractors were negatively or positively correlated we performed the following analyses: for each disease, the ratio between differentially expressed genes in the Th1 attractor and all genes in the Th1 attractor was computed. This analysis was repeated for the Th2 attractor genes. It resulted in a list of ratios for each attractor and for each disease. The Pearson correlation coefficient between those ratios was then computed.

To test if gene pairs in the network model that had counter-regulatory relationships were also negatively correlated, microarray data belonging to healthy controls in each dataset was pooled and Pearson correlation coefficients were calculated for all the gene pairs with counter-regulatory relationships.

Supporting Information

Text S1 References for interactions. In this document we present references supporting interactions introduced in our model network.
Found at: doi:10.1371/journal.pcbi.1001032.s001 (0.12 MB PDF)

Acknowledgements

We thank Dr Michael Langston for valuable comments and for proofreading the manuscript.

Author Contributions

Analyzed the data: FB TC KK. Wrote the paper: MP FB MB. Manual curation of the gene network: MB. Computation of attractors and network analysis: MP. Network extension by data-mining: TC FC EH DS.

References

1. Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 10: 43–55.
2. Reiner SL (2009) Decision making during the conception and career of CD4+ T cells. Nat Rev Immunol 9: 81–82.
3. Cho S, Stanciu LA, Holgate ST, Johnston SL (2005) Increased interleukin-4, interleukin-5, and interferon-gamma in airway CD4+ and CD8+ t cells in atopic asthma. Am J Respir Crit Care Med 171: 224–30.
4. Woodfolk JA (2007) T-cell responses to allergens. J Allergy Clin Immunol 119: 295–6.
5. Gadina M, O'Shea JJ (2009) Immune modulation: Turncoat regulatory t cells. Nat Med 15: 1365.
6. Lund R, Löytömäki M, Naumanen T, Dixon C, Chen Z, et al. (2007) Genome-wide identification of novel genes involved in early Th1 and Th2 cell differentiation. J Immunol 178: 3648–60.
7. Jenssen TK, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28: 21–28.
8. Benson M, Carlsson L, Guillot G, Jernas M, Langston MA, et al. (2006) A network-based analysis of allergen-challenged CD4+ T cells from patients with allergic rhinitis. Genes Immun 7: 514–521.
9. Bosco A, McKenna KL, Devitt CJ, Firth MJ, Sly PD, et al. (2006) Identification of novel Th2-associated genes in T memory responses to allergens. J Immunol 176: 4766–4777.
10. Mendoza L (2006) A network model for the control of the differentiation process in Th cells. BioSystems 84: 101–114.
11. Santoni D, Pedicini M, Castiglione F (2008) Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions. Bioinformatics 24: 1374–1380.
12. Dubrova E, Teslenko M (2010) A SAT-based algorithm for finding attractors in synchronous Boolean networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics In press.
13. Garg A, Xenarios I, Mendoza L, DeMicheli G (2007) An efficient method for dynamic analysis of gene regulatory networks and *in silico* gene perturbation experiments. In: Speed T, Huang H, eds. Research in Computational Molecular Biology, Springer Berlin/Heidelberg, volume 4453 of *Lecture Notes in Computer Science*. pp 62–76.
14. Kauffman SA (1988) Origins of order in evolution: self organization and selection. In: Biomathematics and related computational problems (Naples, 1987). Dordrecht: Kluwer Acad. Publ. pp 311–330.
15. Kauffman S (2004) A proposal for using the ensemble approach to understand genetic regulatory networks. Journal of Theoretical Biology 230: 581–590.
16. Minai-Fleminger Y, Levi-Schaffer F (2009) Mast cells and eosinophils: the two key effector cells in allergic inflammation. Inflamm Res 58: 631–638.
17. Naldi A, Carneiro J, Chaouiya C, Thieffry D (2010) Diversity and plasticity of Th cell types predicted from regulatory network modelling. PLoS Comput Biol 6: e1000912.
18. Jenner RG, Townsend MJ, Jackson I, Sun K, Bouwman RD, et al. (2009) The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. Proc Natl Acad Sci USA 106: 17876–17881.
19. Wang H, Barrenas F, Bruhn S, Mobini R, Benson M (2009) Increased IFN-gamma activity in seasonal allergic rhinitis is decreased by corticosteroid treatment. J Allergy Clin Immunol 124: 1360–1362.
20. Garg A, Di Cara A, Xenarios I, Mendoza L, De Micheli G (2008) Synchronous versus asynchronous modeling of gene regulatory networks. Bioinformatics 24: 1917–1925.
21. Garg A, Mohanram K, Di Cara A, De Micheli G, Xenarios I (2009) Modeling stochasticity and robustness in gene regulatory networks. Bioinformatics 25: i101–109.
22. Chaves M, Sontag ED, Albert R (2006) Methods of robustness analysis for Boolean models of gene control networks. Syst Biol (Stevenage) 153: 154–167.
23. Laubenbacher R, Stigler B (2004) A computational algebra approach to the reverse engineering of gene regulatory networks. J Theoret Biol 229: 523–537.
24. Chaves M, Albert R, Sontag ED (2005) Robustness and fragility of Boolean models for genetic regulatory networks. J Theor Biol 235: 431–449.
25. Garg A, Mendoza L, Xenarios I, DeMicheli G (2007) Modeling of multiple valued gene regulatory networks. Conf Proc IEEE Eng Med Biol Soc. pp 1398–1404.
26. Feinerman O, Veiga J, Dorfman JR, Germain RN, Altan-Bonnet G (2008) Variability and robustness in T cell activation from regulated heterogeneity in protein levels. Science 321: 1081–1084.
27. Mendoza L, Xenarios I (2006) A method for the generation of standardized qualitative dynamical systems of regulatory networks. Theor Biol Med Model 3: 13.

II

Clancy et al. BMC Medical Genomics 2011, 4:28
http://www.biomedcentral.com/1755-8794/4/28

BMC Medical Genomics

RESEARCH ARTICLE Open Access

Immunological network signatures of cancer progression and survival

Trevor Clancy[1*], Marco Pedicini[2], Filippo Castiglione[2], Daniele Santoni[2], Vegard Nygaard[1], Timothy J Lavelle[1], Mikael Benson[3] and Eivind Hovig[1,4,5]

Abstract

Background: The immune contribution to cancer progression is complex and difficult to characterize. For example in tumors, immune gene expression is detected from the combination of normal, tumor and immune cells in the tumor microenvironment. Profiling the immune component of tumors may facilitate the characterization of the poorly understood roles immunity plays in cancer progression. However, the current approaches to analyze the immune component of a tumor rely on incomplete identification of immune factors.

Methods: To facilitate a more comprehensive approach, we created a ranked immunological relevance score for all human genes, developed using a novel strategy that combines text mining and information theory. We used this score to assign an immunological grade to gene expression profiles, and thereby quantify the immunological component of tumors. This immunological relevance score was benchmarked against existing manually curated immune resources as well as high-throughput studies. To further characterize immunological relevance for genes, the relevance score was charted against both the human interactome and cancer information, forming an expanded interactome landscape of tumor immunity. We applied this approach to expression profiles in melanomas, thus identifying and grading their immunological components, followed by identification of their associated protein interactions.

Results: The power of this strategy was demonstrated by the observation of early activation of the adaptive immune response and the diversity of the immune component during melanoma progression. Furthermore, the genome-wide immunological relevance score classified melanoma patient groups, whose immunological grade correlated with clinical features, such as immune phenotypes and survival.

Conclusions: The assignment of a ranked immunological relevance score to all human genes extends the content of existing immune gene resources and enriches our understanding of immune involvement in complex biological networks. The application of this approach to tumor immunity represents an automated systems strategy that quantifies the immunological component in complex disease. In so doing, it stratifies patients according to their immune profiles, which may lead to effective computational prognostic and clinical guides.

Background

Although a link between the immunity and cancer was observed almost 150 years ago [1], the exact nature of the relationship has been developed and debated through several stages of complexity. In recent years, it has been established that the immune system plays crucial roles in tumor development [2], and indeed on patient survival for various cancers [3-7]. Due to a lack of comprehensive analytical approaches, molecular characterization of the roles of the tumor immune component has been somewhat difficult to elucidate on a genome-wide scale.

Current strategies to identify the immune component of tumors tend to employ incomplete manual efforts that do not grade the immune genes. Indeed, even the very definition of an immune gene is unclear, as several interconnected subsystems comprise the totality of immunity. In addition, an analysis of the molecular interactions linked to tumor immunity is usually limited to a pathway-centric paradigm, which is often hindered

* Correspondence: trevor.clancy@rr-research.no
[1]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
Full list of author information is available at the end of the article

by the complexity in which immune pathways are entangled in signaling crosstalk [8]. These challenges are further complicated during cancer progression by the migration of immune cells into unique microenvironments, and by the altered expression of immune genes intrinsic to the tumor. Consequently, as in a tumor gene expression profile, it is not trivial to grade the immune component or identify its related molecular networks.

Multidisciplinary and integrated strategies that handle these and other complex challenges of tumor immunity are increasingly sought after [9-16]. With recent advances in genomics, and increased amounts of latent detailed knowledge in the medical literature, computational approaches can now be developed to study the importance of immune genes and their networks of interactions linked to cancer progression.

Consequently, we have devised a strategy that assigns a ranked immunological relevance score to all human genes for the purpose of profiling the immune component of tumor gene expression. Coupling text mining to information theory, this approach charts immunological relevance onto the human interactome. To apply this strategy in a cancer specific manner, we analyzed melanomas. We first identified immunological signatures that were differentially regulated in the progression from primary stages of skin cancer through to metastases [17]. Survival data from a set of advanced stage melanoma patients were also analyzed, to assess the link between immunological relevance of genes in expression profiles and clinical outcome [5,18].

Our computational approach to assign immunological relevance to genes was benchmarked against manual efforts that identify immune genes, and the strategy was shown to substantiate the performance of existing immunological grading systems. Furthermore, it identified the ranked immunological components of the expression profile of a tumor with its associated networks of interactions. This informative grading of the magnitude of the immune component from patient gene expression profiles may serve as a computational diagnostic and prognostic guide to assess the aggressiveness of a given tumor.

Results

An information theoretical approach to assign immunological relevance to genes

A comprehensive list of 1921 immunology terms was compiled by manual selection of the most relevant terms from the standard biomedical vocabularies of Medical Subject Headings (MeSH) in Medline and the Gene Ontology (GO) controlled vocabulary (see Methods: "Defining the dictionary of terms for immune and neoplasm relevance"). This broad set of terms was collectively considered to be the immunological symbols of communication stored in the over 20 million articles of the biomedical literature (Additional file 1). Using established text mining procedures [19] (see Methods: "Extraction of human genes, immune and neoplasm terms from Medline"), we used these terms and their relationships to gene citations in Medline by capitalizing on the universal feature of coded *information*, present in all forms of communication. By this, it is implied that immune relevant genes have a level of immune information content quantified using this combined set of immune terms in Medline, which is greater than that of genes that play a lesser role in the immune system. Information theory calculations were used to measure the size of the immunological message stored for each human gene with respect to these terms. The probabilities in the information theory calculations were defined through the frequency by which a given gene is cited with a given immune term relative to the number of times the immune term is cited in Medline among all human genes with that term. This measure of immune information content for a gene may be biased by the higher frequency of certain genes being associated overall with the sources of the immune terms, *i.e.* the popularity of a gene among all terms in the biomedical vocabularies. This bias was corrected for using a method in information theory known as the Kullback-Leibler (KL) divergence (see Methods: "An immunological and cancer relevance score for all human genes using information theory and text mining"). The KL score for all human genes was defined as the "immunological relevance" for a gene and termed as such throughout this study (Additional file 2). A similar strategy was also applied to a manual selection of 562 cancer disease terms to determine a genome-wide cancer relevance score for every human gene.

Benchmarking of the immunological relevance score and the extension of immune gene resources

In order to benchmark this immunological relevance for genes, we compared the score against a set of validated immune resources. We utilized gene sets from six manually curated immune efforts (see Methods: "Collating manually curated immune relevant gene sets") that contain independently annotated genes relevant for various aspects of immunity. There were a total of 4833 genes in this integrated set, which had a heterogeneous distribution across the six resources, in that only 82 core immune genes were common to all databases. Many genes in each resource were shared with merely one of the other resources, and few genes were unique to an individual resource (Figure 1). The benchmarking of the immunological relevance score against this set of manually curated immune resources is presented in Figure 2A. The average immunological relevance score

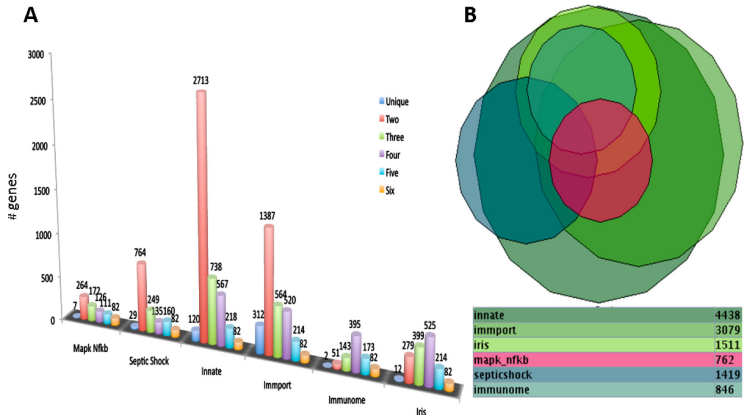


Figure 1 Heterogeneous distribution of genes in immune databases and an incomplete catalogue of immune knowledge. (**A**) Bar chart depicting the shared gene distribution of the immune resources. 82 of the total integrated set of 4833 genes are common to all 6 manually curate resources (orange colored bar). Few genes were unique to an individual database, ranging from a minimum of two for "Immunome" and 122 for the "Innate". (**B**) An approximation using a Venn Euler diagram illustrates the heterogeneous overlap among the different databases. The Innate database being the largest resource has the largest intersections. The septic shock resource has smaller overlaps with the others (with the exception of Innate) highlighting its focus on collating genes related to the response to bacterial toxins during septic shock.

over all genes in each database was determined, compared against each other and the genes not manually curated by these resources. The Immunome [20] ranked the highest among the six manually curated resources in terms of immune information content, reflecting its focus on collating genes enacting functions specific to immune cells. When measuring the immunological relevance of all genes assigned a name by the Human Genome Organization (HUGO) and not catalogued in any of the immune resources, the average approaches zero. The frequency distribution of immunological relevance for all human genes assigned a name in HUGO shows a sharp decline from high to low immunological relevance (Figure 2B), revealing distinct categories of immune and non-immune genes. Moreover, the top ranked genes in the non-curated list represent novel candidates for entry in immune resources (Additional file 3). To assess further the benefit of assigning an automatic immunological relevance score to genes, the integrated set of manually curated genes was compared against two large scale studies that have characterized the human inflammatory response: (1) the *endotoxin response network* from gene expression profiling in human leukocytes [21], and (2) the *inflammation assembly*, which consists of genes detected in genetic variants in inflammatory pathways [22]. The *endotoxin response network* and *inflammation assembly* had 66% and 13% non-overlapping genes with respect to the manually curated resources. The non-correspondence of these six expert resources with large-scale experimental efforts partly indicates the specialized nature of some of these resources and partly may indicate potential in further management of immune knowledge from expert curators. It may also illustrate that there could still be more genes to be implicated in human immunity that are as yet uncharted.

The interactome landscape of immunological and cancer relevance

An affirmed realization from the post genomic era is that no gene functions in isolation, but rather is embedded in a complex network of interacting molecules [23]. Our strategy to profile the immune component of tumors would therefore benefit from an analysis of how immunological relevance relates to the position a gene occupies in complex cellular networks (in this case an integration of three human interactome databases, see Methods: "Constructing a validated human interactome & network analysis"). The creation of a validated and ranked score of a gene's immunological relevance allowed us to chart this score in a landscape setting against cancer relevance and the positional importance (*centrality*) of a gene in the interactome (Figure 3). Centrality is a class of network measurements used to determine the relative importance of a gene in cellular networks. We analyzed five different centrality measures the principal of which being connectivity (i.e. number of interactions per gene). Genes from the six manually curated immune resources on average had a higher connectivity relative the entire interactome (data not shown). Interestingly, increasing

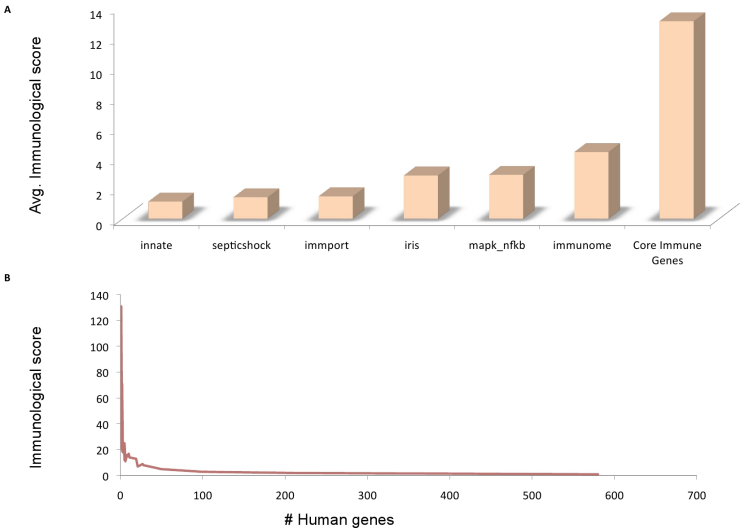


Figure 2 Benchmarking of immunological relevance scores against manually curated immune resources. **(A)** The mean immune score for each database is depicted in the bar chart. The core immune genes are those 82 genes that are common to all immune resources and have a significantly larger amount of information content in comparison to each of the individual immune resources. (**B**) The frequency distribution of all HUGO name assigned genes reveals a sharp decline in immune relevance across the genome.

immunological or cancer relevance showed no strong correlation with connectivity or to any of the four other network centrality measures (Additional file 4). The immune and cancer genes harboring the highest connectivity (network hubs) raise the average, and were unevenly distributed across the heterogeneous interactome landscape (Figure 3). This analysis allowed the detection of scattered peak regions whose genes play driver roles in propagating signals with importance to tumor-immune crosstalk. The classical coordinator of

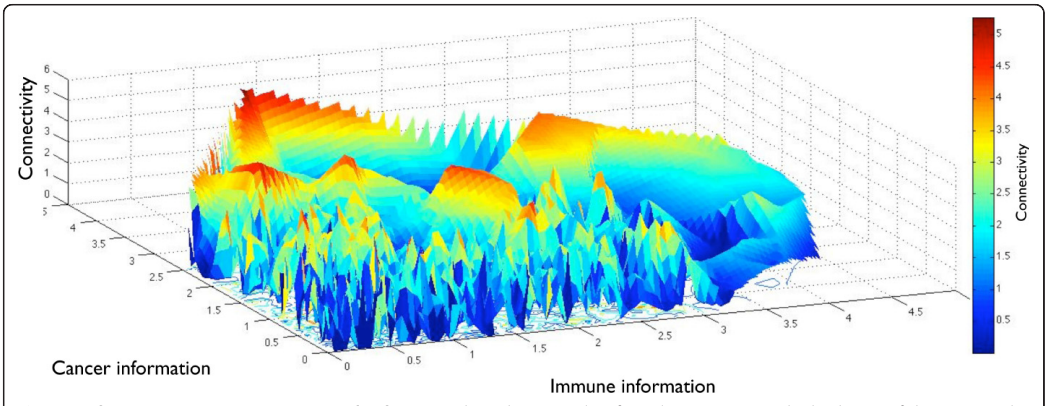


Figure 3 The tumor-immunity interactome landscape. A three-dimensional surface plot representing the landscape of degree centrality (connectivity) of the interactome in the context of immune and cancer relevance: All axes are on the log scale and values above one on the log scale were considered high in terms of immune and cancer relevance. The consideration of one on the log scale as high in terms of immune relevance is made on the basis of the average immune scores for the expert sources ranging from 1 bit and above (see Figure 2). The color scale in the heatmap is representative of the connectivity of each gene in the human interactome. That which is apparent is the distinct areas of scattered high and low connectivity for genes in the cancer-immune landscape. The underlying data for this plot is detailed in Additional File 5.

tumor-immune interactions, *IFNG*, and various T-cell markers were among the highest ranked in this high peak category, as displayed in the underlying interactome landscape data in Additional file 5. We also observed a high degree of correlation (0.75 Spearman's coefficient) between immunological and cancer information content across the genome.

Immunological comparisons of normal tissues and robustness of tissue specific immune interactions

As gene expression profiles of both normal and tumor tissue represent the combined signal of all cell types present in a sample: a global evaluation of the immunological component of normal tissue profiles was attempted, prior to the particular goal of quantifying such for tumors. For this purpose, we calculated the pairwise fold change comparisons of the differentially expressed genes among the 79 tissues profiles from the SymAtlas project [24] and the average immunological relevance score for those differentially expressed genes (shown in heatmap Figure 4A). A gene was considered differentially expressed if it had greater than a two-fold difference in expression between the two tissues under comparison. The pairwise comparisons revealed heterogeneous differences in immunological components among normal tissues (see heatmap in Figure 4A). The

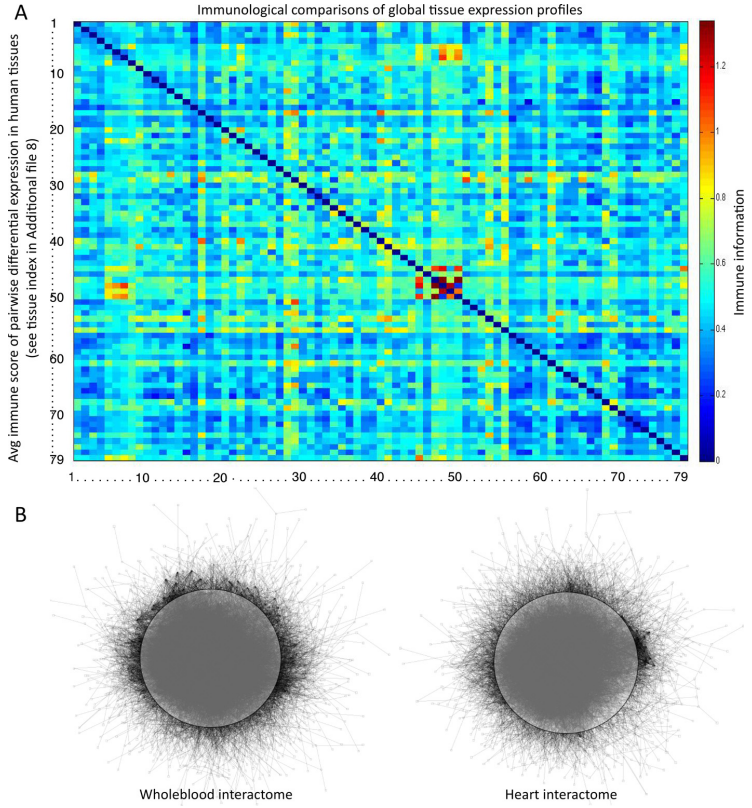


Figure 4 Immunological components of normal tissue. **(A)** Heatmap of the immunological gene expression fold-change comparisons among the 79 tissues from the SymAtlas [24]. This matrix displays the average immune score from those genes that contribute to greater than 2 times fold change difference between each tissue's pairwise comparisons. This combination of expression profiling and immunological grading detects a heterogeneous difference in the immunological components between tissues in a global manner. Both the X and Y-axis are numerical index of the 79 tissues (the mapping of this index to tissue name is listed in Additional File 8). With respect to the robustness of immune genes in the interactome : **(B)** Tissue specific interactome networks for Wholeblood (eccentricity centrality = 0.67) and Heart (eccentricity centrality = 0.72). The difference in the average eccentricity value is only marginally visible by eye as evidenced by a lower symmetry of the Heart network (the same transparent circle drawn on top of the two networks displayed by means of the same algorithm using the software yEd).

comparison, for example, between CD4 and CD8 positive T-cells from the tissue SymAtlas [24] had the largest immunological difference (see heatmap in Figure 4A, columns No 48 and 50 for CD4 and CD8 respectively). The procedure used to determine the differentially expressed genes is detailed in the Methods section entitled: "Microarray gene expression analysis and a composite expression and immunological relevance score".

In order to characterize the differences in the immunological component of these tissues from the perspective of the interactome, we used tissue specific networks previously determined for the SymAtlas tissue profiles [25]. In addition to connectivity, we calculated four other network centrality measures on each of these tissues networks (betweeness, eigenvector, closeness and eccentricity). To test if any of these centrality measures is a discernible property more specific to immune cells, we implemented K-means clustering on all five of the centrality measures across the tissues. Eccentricity was the only measure that classified the tissues in a biological meaningful manner (with K = 9 clusters), in that closely related tissues clustered together (e.g. neurological or immune related tissues, see cluster groups in Additional file 6). Moreover the distribution of the gene eccentricity centralities for each of the tissue interactome networks showed that immune cells had the lowest average eccentricity values (see brown lines, peak value at 0.63 in Figure 5). Leukocytes clustered into three classes, with CD4 and CD8 positive T-cells grouped with wholeblood, lymphoblast precursors into their own separate class, and the remainder of the blood cells profiled (including dendritic cells and NK cells) into a third class (Figure 5 and cluster groups in Additional file 6). The interaction network of a tissue (wholeblood) from the former immune cluster was significantly different from a random network (Wilcox rank, p-val of 0.01) and illustrated graphically in a comparison of this network to that of a non-immune tissue (heart) in Figure 4B. The difference in average eccentricity values between these two tissues is marginally visible in Figure 4B. As immune cells express more immune relevant genes and their eccentricity measures relate to shorter network distance, overall this tissue group clustering suggests that immune genes have more robust connections in the interactome.

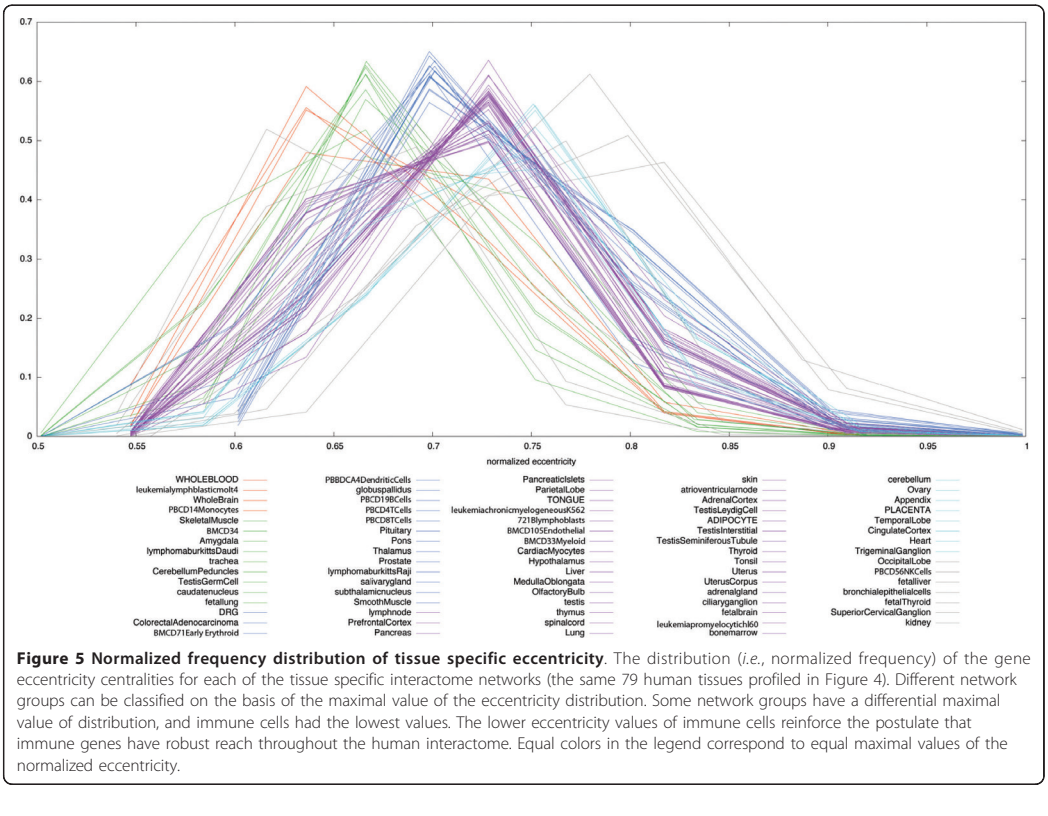


Figure 5 Normalized frequency distribution of tissue specific eccentricity. The distribution (*i.e.*, normalized frequency) of the gene eccentricity centralities for each of the tissue specific interactome networks (the same 79 human tissues profiled in Figure 4). Different network groups can be classified on the basis of the maximal value of the eccentricity distribution. Some network groups have a differential maximal value of distribution, and immune cells had the lowest values. The lower eccentricity values of immune cells reinforce the postulate that immune genes have robust reach throughout the human interactome. Equal colors in the legend correspond to equal maximal values of the normalized eccentricity.

Immunological networks signatures and clinical outcome from expression profiles in melanoma patients

We next extended the principle of tissue expression profiling of immunological signatures in normal tissue to that of expression of normal skin, primary skin tumors and metastatic melanoma [17]. From the pairwise comparison of genes with a greater than two-fold change in expression across these different tissue states, we averaged the immunological score for those genes differentially expressed (> 2 times fold change) and examined how this score differed across the various expression profiles (see Figure 6). Using this average immunological score, we detected clear differences in the stages of progression and related these comparisons to their immune subnetworks from the interactome (see Figure 6). There was a particularly high immunological difference between normal melanocytes and both metastatic and primary melanoma and between normal skin and both metastatic and primary melanoma. The magnitude of the immune component difference between metastatic and normal melanocytes is depicted in Figure 6, along with a related immunological subnetwork of interactions. This network signature shows strong T-cell activation as well as diverse tumor associated chemokine and cytokine activity. There was, however, a much smaller immunological difference between metastatic melanomas and primary melanoma compared to that of normal melanocytes (Figure 6). This suggested that the framework could detect putative signatures of adaptive immunity in mediating transitions at early stages of progression in these patients. The observation that the highest ranked immune genes in these comparisons, *CD4* and *CD8*, were upregulated in primary melanoma and metastasis compared to normal melanocytes signified early and enduring T-cell infiltration. In this comparison, immunological scoring also prioritized markers of innate immune cells such as *PECAM* and *CD14* among others, accompanied by cytokines of inflammatory responses (*IL15, IL7, IL18, IL1A, IL8*). Interestingly, there was also high ranking of an early Th2 tumor-promoting environment demonstrated by presence of the *IL13RA2* gene and the Th1 inhibiting cytokine *IL10*. The smaller amount of immunological information captured in the comparison of primary to metastatic melanoma (Figure 6) was attributable not to high scoring leukocyte or inflammation markers, but by upregulation

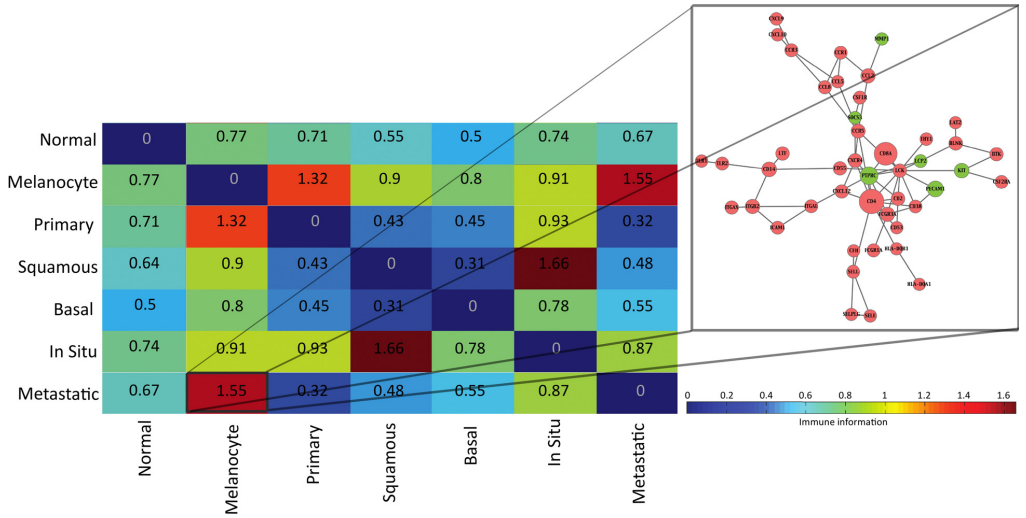


Figure 6 Comparison of the immunological component of skin cancer and states of melanoma progression. A heatmap of the average bits of immune information of the differentially expressed genes (> 2 times fold-change) among the pairwise comparisons of normal skin and skin cancer states. The labels from the left to right columns refer to normal skin tissues: ("Normal"), normal melanocyte ("Melanocyte") and then various states of skin cancer: primary melanoma ("Primary"), squamous cell carcinoma ("Squamous"), basal cell carcinoma ("Basal"), in-situ melanoma ("In Situ") and metastatic melanoma ("Metastatic"). Distinct differences in the immunological component of the various skin cancer and normal states are detected. We have focused here as an example, on the comparison between metastatic melanoma and normal human melanocytes. A subnetwork module from the interactome landscape of those genes with high immunological relevance is displayed. Upregulated genes are color-coded red and downregulated genes are color-coded green in this network. The size of a gene is proportional to the immunological relevance of the gene. There is clearly increased T-cell activity such as the presence of increased expression of *CD8*, *CD4* and *CD3* T-cell markers. This coincides with upregulation of key chemokine and cytokine interactions.

of immunogenic melanoma antigens (*MAGEA2/3*) and downregulation of apoptosis inducing *S100A8/9* cytokines. Summarized gene lists of the top ranked immunological transitions of normal skin, primary and metastatic melanomas are presented in Table 1. In-situ melanoma (MIS) compared to squamous cell carcinoma (SCC) held the highest immunological difference among all the state comparisons (Figure 6). Some of the top ranked immune genes in that comparison included upregulation in SCC relative to MIS of the chemokine *CXCL13* and downregulation of the innate immune gene *LTF*

A composite gene expression and immunological relevance score was used to grade each patient expression profile and find clinical trends to immunological gene signatures (see Methods: "Microarray gene expression analysis and a composite expression and immunological relevance score"). Although the Riker *et al.* study was not accompanied by clinical outcome data, there was a trend in two patients with giant primary melanomas (Breslow thickness of 90 mm) and downregulation of highly relevant immunological genes (p-val, 0.02) compared to 12 other patients with primary melanomas. Using this composite grade, we examined the immunological differences in the outcome, as well as in other clinical features of 57 patients that had reached metastatic melanoma at stage IV [18] and 38 patients at stage III (Bogunovic et at, 2009). Notably, there was a significant association (p-val, 0) with the "high-immune" group of patients as annotated by Jonsson *et al* (as identified by one term, chosen a-priori). Similarly, the strategy detected downregulated highly relevant immunological genes in the patient group that fell into the "proliferative" group of patients (p-val, 0). An upregulated immunological trend was detected in patients that had favorable survival (p-val, 0.1) and was more significant (p-val, 0.02) in those patients categorized with "brisk" immune phenotype (infiltration of CD3 positive lymphocytes). The patient group with NRAS mutations (Q61L) had a correlation with downregulated immunological signatures (p-val, 0.007), hence classifying a group of patients with immune signaling interactions acting downstream of this oncogenic mutation. Patients with hypermethylation of the $p16^{INK4A}$ promoter had trends towards upregulation of genes with high

Table 1 Top ranked immunological transitions of melanoma progression

Gene comparison conditions	Highest graded immune genes	Significance to Melanoma progression
Upregulated (> 2fc) in both primary and metastatic melanoma compared to normal melanocyte (Immunological relevance score for each gene (KL) > 11 bits).	CD4, IL10, CD8A, CD40, IL15, IL7, IL18, TNFSF13B, PTPRC, IL13RA2, IL1A, PECAM1, C5AR1, CD86, ISG20, IL18R1, CD14, ITGB2, ADORA3, FCGR3A, CCL2, IL8, CCR5, FCGR3B	Signatures of T-cell infiltration, T-cell activation and the inflammatory response. Inclusive of the Th1 inhibiting cytokines
Downregulated (> 2fc) in both primary and metastatic melanoma compared to normal melanocyte (Immunological relevance score for each gene (KL) > 0.5 bits).	MME, IL24, DPP4, CYGB, MSC, SLC7A8	Regulation of extracellular matrix (ECM) remodeling, through proteolytic enzymes, and amino acid transporters
Upregulated (> 2fc) in primary melanoma compared to normal melanocyte. Not subject to >2fc in metastasis (Immunological relevance score for each gene (KL) > 2 bits).	IL5, TNF, IL1RN, DARC, HLA-DRB4, CFP, PTPN6, CD1B, ELA2, IL17B, ATP8A2, SLPI, CD27, STAT4, CDA, IL26, DEFB4, NFKBIA, HRH1, XCL1, DEFB1, PDPN, CTSG, SDC1, GATA3, MSMB, CD24, POU1F1, PRDM1, EBF1	Cytokine activity that is pro-survival and towards ECM remodeling. Increased transcriptional activity related to T-cell activation in the primary tumor. Increased presence of MHC class II markers.
Downregulated (> 2fc) in primary melanoma compared to normal melanocyte. Not subject to >2fc in metastasis (Immunological relevance score for each gene (KL) > 1 bit).	BAX, TNFRSF10B, SV2A	Down-regulation is indicative of p53 dysfunction and transduction of apoptosis signals. Overall leading to pro-survival in the primary tumor compared to normal cells
Upregulated (> 2fc) in metastatic melanoma compared to normal melanocyte. Not subject to >2fc in primary. (Immunological relevance score for each gene (KL) > 1 bit).	CCRL2, HLA-DRB1, MDK, C4A, CD55, CD80, FCGR1A, KLRC4, ICAM1, SPI1, HCST, PPBP, FCGR2C, GPR160, CXCL16, FOS, SERPINA1	Mediators of inflammation, angiogenesis, cell growth, and cell migration. Also present are signals of humoral immunity in the form of T-cell activation and B-cell development genes
Downregulated (> 2fc) in metastatic melanoma compared to normal melanocyte. Not subject to >2fc in primary. (Immunological relevance score for each gene (KL) > 1 bit).	KIT, IRF4, MLANA, MMP1	Down regulation of cell adhesion, differentiation factors and regulators of the innate and adaptive immune systems. Possibly promoting the metastatic phenotype
Upregulated (> 2fc) in metastatic melanoma compared to primary (Immunological relevance score for each gene (KL) < 1 bit).	MAGEA3, CSAG2, MAGEA2, GAGE1, MAGEA12, GAGE3, FKBP10	Eliciting immune T cell activation in metastatic tumors, as a consequence of being expressed particularly in the metastatic stages, while having very restricted expression in normal cells
Downregulated (> 2fc) in metastatic melanoma compared to primary (Immunological relevance score for each gene (KL) > 1 bit).	S100A9, S100A8, SLPI, DEFB4, DEFB1, MSMB, CD24, DEFB103A, COL17A1	Altered matrix remodeling and migratory behavior. Dynamic changes in the (ECM) in the metastatic tumors. Inclusive in this is the down regulation of important chemoattractants of innate immune cells

Comparison of progressive melanoma states and their highest weighted immunological relevant genes.

immunological relevance (p-val, 0.05). Overall, the trends with immunological grading of these expression profiles indicated that the assignment of an immunological relevance to genes could classify patient groups with varied immunological signatures. The same analysis was applied to 38 patients from (Bogunovic et al, 2009), and it revealed a significant correlation of upregulated immunological signatures in patients with prolonged survival (p-val, 0.0086) and a significant correlation of downregulated gene with patients that died (p-val, 0.0074). This was also the case in Jonsson et al, where each patient had a unique profile of clinical annotations and immunological gene expression levels (Additional file 7). Interestingly, the authors reported positive correlation with tumor infiltrating leukocytes (TILs) in those patients with favorable survival. A summary of these trends with patient clinical annotations and the immunological profiles for each patient is listed in Additional file 7.

Discussion

The overlap between cancer and immunity has become increasingly well established in recent years. Epidemiologically, 15-20% of cancer deaths are associated to inflammatory conditions [26]. Furthermore, inflammation is a predisposition to cancer, and polymorphisms in cytokine genes are associated to cancer severity [27,28]. Although there is compelling evidence that supports this overlap, an understanding of the molecular mechanisms of what constitutes tumor-immune relationships is far from comprehensive [2,29]. This problem is complicated further by the uniqueness of the microenvironment of each tumor, and the complex interplay between cancer cell immune factors and immune cells infiltrating the tumor.

Gene expression profiling has the potential to provide an improved understanding of these complex relationships and address these challenges. Current approaches to assess the immune component of expression profiles are dependent upon the application of limited predefined sets of immune genes or terms. Prerequisite to the success of manual approaches is the challenge of defining the complete set of immune genes. We have demonstrated that this challenge has not been met. The crux in overcoming this challenge lies in what may be considered to be an immune relevant gene. One option to find immune genes with a role in cancer development is the use of expertly annotated databases [20,30-32]. Our approach improves on the limitations of manual approaches by applying a novel automated procedure that quantifies the immunological relevance for all human genes in bits of information. This score can be directly applied to and provide a more informative and quantitative assessment of the tumor immune component from the gene expression profile. The novel use of information bits to quantify the immunological component may be even further generalized, and applied to any phenotype or any other entity having been assigned symbols of written communication.

Having access to a ranked immunological relevance score for all genes provided an opportunity for analysis of the resulting interactome landscape for tumor immunity. This provided interesting insights into the relationships with levels of immune and cancer information of a gene in the interactome, in light of the new paradigm of network biology [23,33]. These observations in particular add to the debate of the importance of central positions held by cancer [34] and immune [35] genes in the cellular interactome network. Although there is on average higher connectivity for immune and cancer genes in those studies, we illustrated variation about the average, with certain peak genes raising the average connectivity in the interactome landscape.

Tissue specific expression analysis of the immunological relevance score demonstrated that there is a detectable difference among different tissues in the expression of immune genes. Tissue specific network analysis demonstrated that immune genes have distinguishably robust connections within a cells interactome. These observations may be explained by the diverse properties of various tissues to interplay with the immune system in maintaining tissue homeostasis. The strategy of applying a computationally derived immunological score to capture the heterogeneity of the immunological component of normal tissues adds reason to its application as an immunological meta-analysis to cancer transcriptomes. Indeed, quantifying the immunological component of expression studies linked to clinical annotations can lead to informative insights into the immune profiles of patient groups. The necessity and timeliness of applying such a comprehensive computational strategy to tumor expression profiles is highlighted by the increasing reports of immune cell infiltrates in tumor microenvironments as predictors of prognosis and survival in various cancers [4,5,7,36-41].

A proposition for an immunological grading of a tumor based on immune infiltrates has recently been made [42], which would require the expertise of highly trained pathologists. Recent studies in malignant melanoma advocate stratification based on molecular signatures from expression profiling [5,18]. The computational approach described here serves in the automatic identification of ranked immunological signatures and their network of interactions, which leads to a strategy of grading the immunological component of the gene expression of a tumor.

Melanoma was chosen to be the cancer type to demonstrate this strategy, because of the prominent immunological properties of normal skin [43,44] and the strong tendency of melanoma to metastasize [45]. Among the genes harboring some of the highest

immunological relevance, and with expression differences in both primary and metastatic profiles compared to normal skin, were the *CD4* and *CD8* genes. This indicates that our strategy pinpoints possible recruitment of the adaptive immune response at early points in the progression of melanoma in these tumors, which is interesting in the context of increasing evidence that adaptive immunity influences the behavior of human tumors [36]. With respect to melanoma, this further coincides with recent evidence in mice that the metastatic transition is an early event, and that proliferation of disseminating cells is mediated by the function of $CD8^+$ T-cells [46]. Concerning clinical analysis of metastatic melanoma patients, this approach classified the patient group that had immune signatures of upregulated high immunologically relevant genes, and the proliferative-tumor group with down downregulation of high immunologically relevant genes. It was apparent from the clinical analysis that patients had unique combinations of clinical annotations with both up and downregulated genes with high immunological scores. The distinctive immunological profiles for each patient may reflect the uniqueness of the immune component of each microenvironment and the contradictory role immune genes play in regulating cancer development [47].

This strategy does not grade the directionality of these paradoxical roles in the tumor immune response. Rather, it identifies and grades the magnitude of the immune component of the expression profiles. We propose, however, that improving this strategy to do so will precipitate the characterization of detailed mechanisms underlying tumor-immune surveillance, tolerance and escape and facilitate identification of powerful prognostic factors.

Conclusions

We have assigned a ranked immunological relevance score to all human genes applying a novel computational approach that utilizes information theory applied to the medical literature. This score was used to chart immunological relevance against the landscape of protein interaction networks. We propose that this approach can be applied to elucidate the phenotypical component of any complex disease. In this study we focus on tumor immunity and melanoma to demonstrate the ability of this strategy to identify and grade the magnitude of the immune component of patient expression profiles. The capability to analyze tumor transcriptional profiles on a genome-wide scale offers a means to investigate the immunological mechanisms of the complex tumor immune relationships. In so doing, such an approach can classify melanoma patient groups into varied immune profiles that correlate with survival and other clinical phenotypes.

Methods

Defining the dictionary of terms for immune and neoplasm relevance

By doing manual searches in the Gene Ontology (GO) [48] and the Medical Subject Headings (MeSH) (http://www.nlm.nih.gov/mesh/) resources and documenting those terms deemed relevant for the context, we compiled a list of 1921 immune and 562 neoplasm context terms. This resulted in a comprehensive term list from structured vocabularies that define the contexts in our analysis. The manual searches were implemented using domain knowledge of immunity and cancer. Strict scrutiny of relevance to the context was applied before acceptance of a term into the context term list. The manual searches in MesH and GO produced a candidate list of terms. Each candidate term was read and then categorized as being relevant or not relevant for immunity or cancer based on the expert knowledge of an immunologist or cancer researcher, respectively. As the purpose of this study was to quantify the size of the immune component of tumor samples, a broad scope of immune terms was accepted, each term has an association of an immune function, process, cellular anatomy or immune condition according to the scrutiny of the immunologist. The complete list of chosen immune and neoplasm terms is presented in Additional file 1.

Extraction of human genes, immune and neoplasm terms from Medline

One of the important elements in the approach is to identify literature co-citations of human genes and their associated GO and MeSH terms by using an established method in text mining [19]. Here is a brief summary of this method with more detail in the referenced article. All official symbols, names and alias symbols for human genes compiled from the Human Genome Organization (HUGO) (http://www.genenames.org/), OMIM (http://www.ncbi.nlm.nih.gov/omim/), and EntrezGene (http://www.ncbi.nlm.nih.gov/gene), were automatically extracted from all Medline article titles and abstracts. The genes are indexed to PubMed IDs after a natural language processing (NLP) step of the Medline abstracts that involves procedures in part of speech tagging (POS) and noun chunking, the purpose of which is to remove false positives of biological term mentions. Some other steps in obtaining the gene citation data of higher quality is to remove abbreviation type false positives, which occur frequently because gene symbols often coincide with other abbreviations having no connection or relevancy with the gene symbol. Such data quality steps yield a greater number unambiguous gene symbol citations in text with an improved precision. In a similar manner to the extractions of gene from Medline text GO terms are extracted using NLP techniques of POS

and the GO terms mapped to their corresponding identifiers and indexed against noun chunks in Medline sentences. MeSH terms are indexed to Medline abstracts by using the National Library of Medicine's (NLM) annotations of MeSH terms to articles.

An immunological and cancer relevance score for all human genes using information theory and text mining

The principle of Shannon's entropy was first tested as a sensible measure of information content applied to gene associations derived from text mining. This was further refined using the Kullback-Leibler (KL) score, thus correcting for bias introduced by the popularity of the gene to be co-cited in all of Medline which we found to inherent in the Shannon entropy calculations.

In these information theory aproach we interpreted gene co-citation events in medical articles with terms from a lexicon of expertly chosen annotations from a context as an information coding system for the context (the context of immunity and cancer in this study). The frequency of co-citation events of a manually annotated context term i extracted from Medline abstracts using text mining and co-cited with a human gene x is treated as a message. This message is detected within each element of an alphabet of symbols of size N, where N is the total number of annotated terms in the lexicon of that context. Immune and cancer experts manually chose the elements of the alphabet of symbols N from the structured biological vocabularies of GO and MeSH. Thus we view the literature association between a gene and a context term as the observance of a symbol describing an element of that contextual message and the probability of that event occurring is:

$$p(x) = \frac{g_i}{iT_g}$$

g_i is the number of co-citations for a gene with a context term i in Medline and iT_g is the total number of times the context term i is cited with all human genes in all of Medline (the total gene co-citation space of the context term). Hence, the continuously expanding 20 million articles of Medline is the source emitting these symbols with probabilities (p1, p2, . . ., pN) and these are the symbols of communication that define an immunological (or other contextual) score for all human genes. We assume that the symbols are emitted independently for each gene. In this assumption the probability that a gene is associated to, for example, the immune term "T-cell differentiation" in Medline is independent of its association to the immune term "Macrophage" and their probabilities are computed independently. These probabilities of events (p1, p2, . . ., pN) give discrete values that can be used to detect the size of a message and thus the contextual information content for each gene as defined by Shannon's entropy [49]

$$Hc = \sum_{i}^{N} p(x)\log_2 p(x)$$

Although the Shannon entropy provided the accurate size of the information content for each gene, it did not account for bias introduced by the popularity of the gene Medline. We therefore refined the information theory approach to correct for this bias. This bias was defined as the popularity of the gene to be co-cited in all of Medline, *i.e.* its probability of co-citation among all GO and MeSH terms in the gene co-citation space of Medline. We quantified this bias and corrected for it using the Kullback-Leibler (KL or "relative entropy") calculation to create a more accurate measure of information content that can be used as the immunological and cancer score for each gene. The KL was used to determine the divergence of the observed probability $p(x)$, described above, from an assumed incorrect distribution, which we take as the popularity of the gene in the total Medline gene co-citation space $q(x)$:

$$q(x) = \frac{g_T}{GS_T}$$

Where g_T is the number of co-citations for a gene with all GO and MeSH terms in Medline and GS_T is the total number of co-citation events for all GO and MeSH terms with all human genes in all of Medline (the total gene co-citation space of GO and MeSH, the source of the immune and cancer context terms chosen by domain experts). This measures the expected amount of information required to code a message from a context term for a gene $p(x)$ when using a code based on the assumed incorrect probability $q(x)$ rather than using a code based on $p(x)$ and is defined by Kulback-Leibler KL as:

$$KL = \sum_{i}^{N} p(x)\log_2 \frac{p(x)}{q(x)}$$

As this relative entropy score (KL) corrects for the bias $q(x)$ for each gene, it was used as to calculate the "immunological and cancer relevance" score throughout this study.

Collating manually curated immune relevant gene sets

The immunology gene sets were compiled from the following manually curated sources: (1) Immport (https://www.immport.org), (2) Immunome [20], (3) Iris [31], (4) Mapk-Nfkb (ref), (5) Septic Shock (http://www.septicshock.org) and (6) InnateDB [30]. The HUGO (http://

www.genenames.org/) symbol for genes provides a unique identifier for human genes and is ideal for the integration of text mining derived knowledge. It was used in this study to integrate and determine the overlapping descriptive statistics for each of the six databases and visualized in Venn diagrams in the VennMaster software [50] to approximate their intersections by incorporating the gene set size information. Similarly genes from two efforts to catalogue the inflammatory response [21,22] were integrated using their HUGO gene symbols and compared to the unified immune gene set from the above six different sources mentioned above.

Constructing a validated human interactome & network analysis

We constructed a human gene network by integrating binary human interactions from IntAct [51], BioGRID [51] and HPRD [52]. Each of these datasets of binary interacting protein pairs was downloaded from their source and the unique ids of the interactors were cross-referenced to their NCBI gene IDs and official Gene Symbols. This resulted in a unified set of binary NCBI gene ID interactor pairs, with their corresponding official gene symbols. The interaction data was limited to these sources as they consist of validated protein-protein interactions with experimental evidence curated from critical reading of the scientific literature by expert biologists.

This integrated data set is represented as an undirected, unweighted network, where G = (V,E) comprising of a set of nodes V and edges E. Each node represents a human gene and each edge represents a pair of genes (u,v) as a representation of a binary interaction in the human interactome. If there exists a physical binary interaction between u and v, in at least one of the protein products of each gene, an edge is connected. The tissue specific interactomes were derived from the entries in the three protein interaction databases mentioned above and the tissue expression annotations from in a recent study integrating tissue specific interactions from 79 human tissues [25].

Network centrality analysis was carried out on the networks by means of calculating five measures of centrality for each gene in the interactome (Connectivity, betweeness, eccentricity, closeness and eigenvector). A descriptions of equations implemented for these measures and full details of their context to protein networks in cancer are summarized here [53]

Microarray gene expression analysis and a composite expression and immunological relevance score

Tissue specific gene expression data from the Symatlas project [24] was analyzed to detect pairwise differential expression across the 79 specific tissues [25] (Additional file 8). We considered a gene differentially expressed between any pair of tissues and therefore viable for further analysis if there was greater than a two times fold-change in expression. The average immunological score was then determined for these differentially expressed genes across all tissue pairs. A similar approach was used for profiles in the progressive states of skin cancer. For the gene expression profile linked to patient survival probes [18] strict criteria were applied to reduce false positive signals in that only those probes with detection p-value < 0.01 in more than 50 out of the total 57 patients were used. The software used to calculate the detection p-values (Illumina BeadStudio) uses a nonparametric method for the computation of detection p-values. In this method the z-values of the probe signals are ranked relative to the z-values of the negative control signals. These were quantile normalized [54], and log2 transformed. Each probe signal intensity measurement (S) was given a fold change relative to that probes mean signal intensity (MSI) across all patients (P) and utilized to create a weighed composite signal intensity and immunological score for each gene (W_g):

$$\sum_{1}^{P} \log_2\left(\frac{S}{MSI}\right).KL$$

The weighted expression and immune score for each gene was then summated across all genes (M) for each patient to generated a weighted immune score for each patient (W_p):

$$\sum_{1}^{M} W_g$$

The patient scores where compared to the clinical annotations to find correlations between the weighted immunological score (W_p) and the clinical phenotypes. Monte Carlo simulations with 10,000 draws were used to create a null distribution for each comparison. For numerical phenotypes Pearson's correlation were used.

Additional material

Additional file 1: Tables of manually curated immune and neoplasia terms. These are the terms used to define the dictionary of terms for immune and neoplasm context. Manually selected from GO and MeSH using domain knowledge.

Additional file 2: Genome-wide ranked Immunological and neoplasia relevance score for genes. Table depicting the immunological and cancer relevance score for all human genes quantified in bits using information theory calculation with Kullback-Leibler adjustments

Additional file 3: Immunological relevance of non-curated genes. Ranked immunological relevance of genes not populated in the manually curated immune gene resources

Additional file 4: Relationship of both immunological and cancer relevance to network centrality. Tables reporting the Pearson correlation coefficients of immunological and cancer relevance to the principle network centrality measures of the human interactome.

Additional file 5: Tumor immunity interactome landscape. The underlying data behind Figure 3, quantifying in bits of information the immunological and cancer relevance charted against connectivity in the interacome

Additional file 6: Table of the k-means classification by means of the eccentricity centrality measure, showing biologically meaningful classes of tissues. K-means classification of tissue groups shown in Figure 5 (parameter K = 9). Determined by means of the eccentricity centrality measure for each of the tissue specific interactomes from the SymAtlas [24].

Additional file 7: Bogunovic et al, 2009 distinct patient profiles and relationship to clinical phenotypes. Composite expression and immunological relevance score for all genes in each patient in this study. Demonstrated here as an example to offer an overview of the diversity and uniqueness of the immunological profile, detected by this approach in each individual patient samples.

Additional file 8: Normal tissue index. Index for the 79 normal tissues from the SymAtlas [24] depicted in the heatmap of immunological comparisons in Figure 4A

Acknowledgements
The research leading to these results has received funding from the European Commision (FP6-2005-NEST-PATH, No. 043241 - ComplexDis and FP7-2008, No 223367-MultiMod).

Author details
[1]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway. [2]Institute for Computing Applications, National Research Council, Rome, Italy. [3]The Unit for Clinical Systems Biology, University of Gothenburg, Gothenburg, Sweden. [4]Institute of Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway. [5]Department of Informatics, The University of Oslo, Oslo, Norway.

Authors' contributions
TC conceived and developed the information theory approach to quantify the immune and cancer components. TC, TJL, and EH designed and planned the study. MP, DS, FC and TC performed the network analysis. VN performed the gene expression profiling. EH, TJL, MB, and TC applied and developed the manual curation pipeline and biological interpretation of the data. TC drafted the manuscript. EH, TC and TJL wrote the final manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 2 September 2010 Accepted: 31 March 2011
Published: 31 March 2011

References

1. Virchow RL: In *Die Krankhaften Geschwülste Dreissig Vorlesungen gehalten während des Wintersemesters 1862-63 Vierte Vorlesung. Volume 1*. Berlin; 1863:65.
2. Mantovani A, Allavena P, Sica A, Balkwill F: **Cancer-related inflammation.** *Nature* 2008, **454**:436-444.
3. Galon J: **Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome.** *Science* 2006, **313**:1960-1964.
4. Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, Fisher RI, Braziel RM, Rimsza LM, Grogan TM, *et al*: **Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells.** *N Engl J Med* 2004, **351**:2159-2169.
5. Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu YL, Adams S, Darvishian F, Berman R, Shapiro R, Pavlick AC, *et al*: **Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival.** *Proc Natl Acad Sci USA* 2009, **106**:20429-20434.
6. Disis ML: **Immune Regulation of Cancer.** *Journal of Clinical Oncology* 2010.
7. Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, Makrigiannakis A, Gray H, Schlienger K, Liebman MN, *et al*: **Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer.** *N Engl J Med* 2003, **348**:203-213.
8. Fraser IDC, Germain RN: **Navigating the network: signaling cross-talk in hematopoietic cells.** *Nat Immunol* 2009, **10**:327-331.
9. Mlecnik B, Tosolini M, Charoentong P, Kirilovsky A, Bindea G, Berger A, Camus M, Gillard M, Bruneval P, Fridman WH, *et al*: **Biomolecular network reconstruction identifies T-cell homing factors associated with survival in colorectal cancer.** *Gastroenterology* 2010, **138(4)**:1429-1440.
10. Davis MM: **A prescription for human immunology.** *Immunity* 2008, **29**:835-838.
11. Strausberg RL: **Tumor microenvironments, the immune system and cancer survival.** *Genome Biol* 2005, **6**:211.
12. Wang E, Panelli MC, Monsurró V, Marincola FM: **A global approach to tumor immunology.** *Cell Mol Immunol* 2004, **1**:256-265.
13. Petrovsky N, Brusic V: **Computational immunology: The coming of age.** *Immunol Cell Biol* 2002, **80**:248-254.
14. Mlecnik B, Sanchez-Cabo F, Charoentong P, Bindea G, Pagès F, Berger A, Galon J, Trajanoski Z: **Data integration and exploration for the identification of molecular mechanisms in tumor-immune cells interaction.** *BMC Genomics* 2010, **11(Suppl 1)**:S7.
15. Bindea G, Mlecnik B, Fridman WH, Pages F, Galon J: **Natural immunity to cancer in humans.** *Curr Opin Immunol* 2010, **22(2)**:215-222.
16. Bedognetti D, Wang E, Sertoli MR, Marincola FM: **Gene-expression profiling in vaccine therapy and immunotherapy for cancer.** *Expert Rev Vaccines* 2010, **9(6)**:555-565.
17. Riker AI, Enkemann SA, Fodstad O, Liu S, Ren S, Morris C, Xi Y, Howell P, Metge B, Samant RS, *et al*: **The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis.** *BMC Med Genomics* 2008, **1**:13.
18. Jonsson GB, Busch C, Knappskog S, Geisler J, Miletic H, Ringner M, Lillehaug JR, Borg A, Lonning PE: **Gene Expression Profiling-Based Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome.** *Clin Cancer Res* 2010.
19. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28**:21-28.
20. Ortutay C, Vihinen M: **Immunome Knowledge Base (IKB): An integrated service for immunome research.** *BMC Immunol* 2009, **10**:3.
21. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, *et al*: **A network-based analysis of systemic inflammation in humans.** *Nature* 2005, **437**:1032-1037.
22. Loza MJ, Mccall CE, Li L, Isaacs WB, Xu J, Chang BL: **Assembly of inflammation-related genes for pathway-focused genetic analysis.** *PLoS ONE* 2007, **2**:e1035.
23. Barabasi AL: **Scale-Free Networks: A Decade and Beyond.** *Science* 2009, **325**:412-413.
24. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16)**:6062-6067.
25. Bossi A, Lehner B: **Tissue specificity and the human protein interaction network.** *Mol Syst Biol* 2009, **5**:260.
26. Newton R: **Infections and human cancer.** *Ann Oncol* 2000, **11(9)**:1081-1082.
27. Balkwill F, Mantovani A: **Inflammation and cancer: back to Virchow?** *Lancet* 2001, **357**:539-545.
28. Balkwill F, Charles KA, Mantovani A: **Smoldering and polarized inflammation in the initiation and promotion of malignant disease.** *Cancer Cell* 2005, **7**:211-217.
29. Porta C, Larghi P, Rimoldi M, Grazia Totaro M, Allavena P, Mantovani A, Sica A: **Cellular and molecular pathways linking inflammation and cancer.** *Immunobiology* 2009, **214**:761-777.
30. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan THW, Shah N, *et al*: **InnateDB: facilitating systems-level analyses of the mammalian innate immune response.** *Mol Syst Biol* 2008, **4**:1-11.
31. Kelley J, de Bono B, Trowsdale J: **IRIS: a database surveying known human immune system genes.** *Genomics* 2005, **85**:503-511.

32. Lynn DJ, Chan C, Naseer M, Yau M, Lo R, Sribnaia A, Ring G, Que J, Wee K, Winsor GL, *et al*: **Curating the innate immunity interactome.** *BMC Syst Biol* 2010, **4(1)**:117.
33. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101-113.
34. Jonsson PF: **Global topological features of cancer proteins in the human interactome.** *Bioinformatics* 2006, **22**:2291-2297.
35. Ortutay C, Vihinen M: **Efficiency of the immunome protein interaction network increases during evolution.** *Immunome Res* 2008, **4**:4.
36. Hodi FS, Dranoff G: **The biologic importance of tumor-infiltrating lymphocytes.** *J Cutan Pathol* 2010, **37(Suppl 1)**:48-53.
37. Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, Mlecnik B, Kirilovsky A, Nilsson M, Damotte D, *et al*: **Effector memory T cells, early metastasis, and survival in colorectal cancer.** *N Engl J Med* 2005, **353**:2654-2666.
38. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, *et al*: **Type, density, and location of immune cells within human colorectal tumors predict clinical outcome.** *Science* 2006, **313**:1960-1964.
39. Piras F, Colombari R, Minerba L, Murtas D, Floris C, Maxia C, Corbu A, Perra MT, Sirigu P: **The predictive value of CD8, CD4, CD68, and human leukocyte antigen-D-related cells in the prognosis of cutaneous malignant melanoma with vertical growth phase.** *Cancer* 2005, **104**:1246-1254.
40. van Houdt IS, Sluijter BJR, Moesbergen LM, Vos WM, de Gruijl TD, Molenkamp BG, van den Eertwegh AJM, Hooijberg E, van Leeuwen PAM, Meijer CJLM, *et al*: **Favorable outcome in clinically stage II melanoma patients is associated with the presence of activated tumor infiltrating T-lymphocytes and preserved MHC class I antigen expression.** *Int J Cancer* 2008, **123**:609-615.
41. Sato E, Olson SH, Ahn J, Bundy B, Nishikawa H, Qian F, Jungbluth AA, Frosina D, Gnjatic S, Ambrosone C, *et al*: **Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer.** *Proc Natl Acad Sci USA* 2005, **102**:18538-18543.
42. Pagès F, Galon J, Dieu-Nosjean MC, Tartour E, Sautès-Fridman C, Fridman WH: **Immune infiltration in human tumors: a prognostic factor that should not be ignored.** *Oncogene* 2010, **29**:1093-1102.
43. Kupper TS, Fuhlbrigge RC: **Immune surveillance in the skin: mechanisms and clinical consequences.** *Nat Rev Immunol* 2004, **4**:211-222.
44. Nestle FO, Di Meglio P, Qin JZ, Nickoloff BJ: **Skin immune sentinels in health and disease.** *Nat Rev Immunol* 2009, **9**:679-691.
45. Chin L, Garraway LA, Fisher DE: **Malignant melanoma: genetics and therapeutics in the genomic era.** *Genes & Development* 2006, **20**:2149-2182.
46. Eyles J, Puaux AL, Wang X, Toh B, Prakash C, Hong M, Tan TG, Zheng L, Ong LC, Jin Y, *et al*: **Tumor cells disseminate early, but immunosurveillance limits metastatic outgrowth, in a mouse model of melanoma.** *J Clin Invest* 2010, **120**:2030-2039.
47. De Visser KE, Eichten A, Coussens LM: **Paradoxical roles of the immune system during cancer development.** *Nat Rev Cancer* 2006, **6**:24-37.
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
49. Shannon CE: **A Mathematical Theory of Communication.** *The Bell System Technical Journal* 1948, 379-423, 623-656.
50. Kestler HA, Müller A, Gress TM, Buchholz M: **Generalized Venn diagrams: a new method of visualizing complex genetic set relations.** *Bioinformatics* 2005, **21**:1592-1595.
51. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, *et al*: **IntAct–open source resource for molecular interaction data.** *Nucleic Acids Research* 2007, **35**:D561-565.
52. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TKB, Chandrika KN, Deshpande N, Suresh S, *et al*: **Human protein reference database as a discovery resource for proteomics.** *Nucleic Acids Research* 2004, **32**:D497-501.
53. Platzer A, Perco P, Lukas A, Mayer B: **Characterization of protein-interaction networks in tumors.** *BMC Bioinformatics* 2007, **8**:224.
54. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**, Article3.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1755-8794/4/28/prepub

doi:10.1186/1755-8794-4-28
Cite this article as: Clancy *et al.*: Immunological network signatures of cancer progression and survival. *BMC Medical Genomics* 2011 **4**:28.

III